SYSTEM APPROACH TO ROBUST ACOUSTIC ECHO CANCELLATION THROUGH SEMI-BLIND SOURCE SEPARATION BASED ON INDEPENDENT COMPONENT ANALYSIS

A Thesis
Presented to
The Academic Faculty

by

Ted S. Wada

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
August 2012

SYSTEM APPROACH TO ROBUST ACOUSTIC ECHO CANCELLATION THROUGH SEMI-BLIND SOURCE SEPARATION BASED ON INDEPENDENT COMPONENT ANALYSIS

Approved by:

Professor Biing-Hwang (Fred) Juang,
Advisor
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Professor David V. Anderson
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Professor Xiaoli Ma
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Professor Jeff S. Shamma
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Professor Brani Vidakovic
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Date Approved: 22 June 2012

To my family, friends, and all the acquaintances
I have come in contact during my life's random walk.

ACKNOWLEDGEMENTS

Many thanks go out to countless people that I met throughout the academic journey at Georgia Tech that has taken a large part of my life so far.

To start with, I want to recognize the fellow research group members and graduate students from the Center for Signal and Image Processing (CSIP) and the school of Electrical and Computer Engineering (ECE): Antonio Moreno-Daniel, Rungsun Munkong, Woojay Jeon, Gaofeng Yue, Qiang Fu, Enrique Robledo-Arnuncio, Dwi Sianto Mansjur, Soohyun Bae, Yong Zhao, Sunghwan Shin, Umair Altaf, Jason Wung, Mingyu Chen, Chao Weng, Jin Wang, Yu Tsao, Yeongseon Lee, Byungki Byun, Jinwoo Kang, Gregory Krudysz, Kaustubh Kalgaonkar, Toygar Akgun, Kun Shi, Thao Tran, Ilseo Kim, Jonathan Kim, Xusheng Sun, Deryck Yeung, Byungchil Kim, and Calvin Xu. I apologize to those I may have left out.

I would like to acknowledge the visiting scholars and professors I have come across during my long-term stay at Georgia Tech: Drs. Marco Siniscalchi, Shigeki Miyabe, Francesco Nesta, Taras Butko, Taeyoon Kim, Shinji Watanabe, Tomoko Matusi, Tomohiro Nakatani, Hiroshi Saruwatari, and Thrasyvoulos Pappas. I especially thank Dr. Miyabe and Dr. Nesta, without whom I would not have had a direct exposure to the topic of blind source separation that became an important aspect of my dissertation.

I would like to direct my appreciation to the professors who agreed to share their precious time to be on my dissertation committee: Profs. David Anderson, Xiaoli Ma, Jeff Shamma, and Brani Vidakovic. I also would like to thank those behind the scene who kept or still keep ECE and CSIP running: Prof. David Hertling, Prof. Bonnie Ferri, Prof. Douglas Williams, Marilou Mycko, Daniela Staiculescu, Suzzette Willingham, Tasha Torrence, Jacqueline Trappier, Christopher Malbrue, Christina Bourgeois, Chris McGahey, Diana Fouts, Pat Dixon, Christy Ellis, Tammy Scott, Jennifer Lunsford, Lisa Gardner, and Stacie Speights.

I am forever grateful to the bosses, mentors, and co-workers at the companies I interned

with for giving me the opportunities to gain valuable work experiences: Drs. Eric Diethorn and Gary Elko at Avaya, Inc., Drs. Rick Younce and Rafid Sukkar at Tellabs, Inc., Drs. Qi Li and Manli Zhu at Li Creative Technologies, Inc., and Drs. Juin-Hwey Chen and Jes Thyssen at Broadcom, Inc.

Most importantly, I would like to reflect the sincerest respect and gratitude towards my advisor, Prof. Biing-Hwang Juang. I could not have stuck around for so long at Georgia Tech without his patience and generosity. Dr. Juang's critical, insightful, and practical view on scientific research and everyday life should continue to guide my future endeavors.

Finally but not least, I would like to thank my mother Junjitsu and father Motonaka for giving me an identity that can never be taken away, and my brothers David and Roy for ensuring that I am not alone in this world.

TABLE OF CONTENTS

DEDICATION iii

ACKNOWLEDGEMENTS iv

LIST OF TABLES ix

LIST OF FIGURES x

SUMMARY xiv

I INTRODUCTION 1

1.1 Motivations 1

1.2 Objectives 4

1.3 Outline 6

II PROBLEM BACKGROUND 9

2.1 Acoustic Echo Cancellation (AEC) 9

2.1.1 Mean Square Error (MSE) Minimization 9

2.1.2 Least Mean Square (LMS) Algorithm 11

2.1.3 Orthogonality Principle 12

2.1.4 Performance Measures 14

2.2 Multi-channel AEC (MCAEC) 15

2.2.1 Single-channel Solution to MCAEC 15

2.2.2 Non-uniqueness Problem 16

2.2.3 Coherence Measure 18

2.3 Source Separation 19

2.3.1 Blind Source Separation (BSS) 19

2.3.2 Independence Maximization via Independent Component Analysis (ICA) 20

2.3.3 Semi-Blind Source Separation (SBSS) 22

2.4 Effect of Distortions on AEC and BSS 25

2.4.1 Linear Distortion 25

2.4.2 Nonlinear Distortion 26

2.4.3 Sampling Rate Mismatch 28

III ROBUST AEC VIA RESIDUAL ECHO ENHANCEMENT (REE) 30
3.1 Conventional Approaches 33
3.1.1 Robustness against Non-stationarity 33
3.1.2 Robustness against Ambient Noise 33
3.1.3 Robustness against Double-Talk 34
3.2 Residual Echo Enhancement 35
3.2.1 Error Recovery Nonlinearity (ERN) 35
3.2.2 Connection to Conventional Approaches 42
3.2.3 Analysis of ERN 45
3.2.4 Connection between LMS, ICA, and SBSS 47
3.2.5 Block-Iterative Adaptation (BIA) 50
3.3 Experimental Evaluation 52
3.3.1 Time-Domain AEC with Ambient Noise 53
3.3.2 Time-Domain AEC with Speech Coding Distortion 55
3.3.3 Frequency-Domain AEC with Ambient Noise and Double-Talk . . 59
IV SYSTEM PERSPECTIVE OF DECORRELATION FOR AEC 64
4.1 Inter-channel Decorrelation by Resampling 67
4.1.1 Review of the Non-uniqueness Problem 67
4.1.2 Systematic Link between BIA, Decorrelation, and Resampling . . 68
4.1.3 Decorrelation by Resampling (DBR) 69
4.1.4 Decorrelation via Frequency-Domain Resampling (FDR) 73
4.1.5 Decorrelation via Sub-band Resampling (SBR) 76
4.2 Extension of Robust AEC System 77
4.2.1 Robust MCAEC 77
4.2.2 Variations in BIA 77
4.2.3 Multi-delay Filter (MDF) 78
4.3 Experimental Evaluation 80
4.3.1 Robust MCAEC with and without DBR 80
4.3.2 Robust MCAEC with FDR 86
4.3.3 Robust MCAEC with SBR 92
4.3.4 Robust MDF 96

V ROBUST MCAEC VIA SBSS 98
5.1 Generalization of MCAEC by SBSS 100
5.1.1 SBSS Model 100
5.1.2 Non-uniqueness Problem in SBSS 103
5.1.3 Derivation of Steady-State Solution 104
5.1.4 Connection between MSE and ICA 108
5.1.5 Constraint on Separation Matrix 110
5.2 Algorithm Design and Related Issues 115
5.2.1 Online Implementation of SBSS 115
5.2.2 Proposed SBSS Algorithm 119
5.3 Experimental Evaluation 121
5.3.1 Evaluation Methods 121
5.3.2 Experimental Results 122
VI CONCLUSIONS 140
6.1 Contributions 143
6.2 Future Research Suggestions 144
APPENDIX A BAYESIAN ESTIMATION 145
APPENDIX B FISHER INFORMATION 147
APPENDIX C GENERALIZED DECORRELATION 148
REFERENCES 149
VITA 159

LIST OF TABLES

1 Three sample-based adaptive algorithms obtained through least-square optimization and commonly used for AEC: least mean square (LMS), affine projection (AP), and recursive least squares (RLS). $E\{\cdot\}$ is the expectation operator, $e(n)$ is the estimation error, L is the filter order, P is the AP order, and $0 < \lambda \leq 1$ is the forgetting factor. 10

2 Sampling rate mismatch, measured in terms of parts per million (ppm), for six portable audio capturing devices with the nominal sampling rate of 16 kHz. 29

3 tERLE comparison (dB, higher is better). 88

4 SSRR comparison (dB, higher is better). 88

5 LSD comparison (lower is better). 88

6 Average tERLE (dB, left column) and misalignment(dB, right column). . . 90

7 Processed speech quality comparison. 94

8 Average tERLE (dB, higher is better). 95

9 Average misalignment (dB, lower is better). 95

10 Average tERLE (dB, higher is better). 96

11 Average misalignment (dB, lower is better). 96

12 Average tERLE (dB, left column) and misalignment(dB, right column). . . 97

13 Outline of the proposed SBSS algorithm. 120

14 Procedure for adaptation of the de-mixing matrix. 120

15 Summary of parameters used during adaptation. 123

LIST OF FIGURES

1 Model for acoustic echo inside the loudspeaker-enclosure-microphone system (LEMS). A loudspeaker power amplifier (PA) is often modeled by a memoryless saturation nonlinearity that distorts the far-end (reference) signal $x(n)$, whereas the effect of a microphone pre-amp is usually ignored. Impulse responses $h_d(n)$, due to a direct echo path, and $h_r(n)$, due to reverberation (*i.e.*, reflections), are combined together as a single room impulse response (RIR) $h(n)$. A local noise $v(n)$ is added directly to the acoustic echo $d(n)$, and the resulting near-end signal $y(n)$ is sent back to the far end. 2

2 Model for acoustic echo cancellation (AEC). There may be linear or nonlinear distortion applied to the reference signal $\bar{x}(n)$ or the acoustic echo $\bar{d}(n)$ that can prevent a linear finite impulse response (FIR) adaptive filter from properly identifying the RIR $h(n)$. Any remaining echo component in the estimation error $e(n)$ is reduced further through residual echo suppression (RES). 2

3 Implementation of AEC in the telecommunications network. Speech coding distortion on the acoustic echo that severely degrades the AEC performance is unavoidable in such a scenario. 4

4 Effect of additive noise v on the least mean square (LMS) estimation of the optimal filter coefficient vector $\mathbf{w}_{opt}$. The optimal solution is obtained as long as $E\{vx\} = 0$, *i.e.*, $E\{e\} = E\{(\bar{e} + v)\mathbf{x}\} = E\{\bar{e}x\}$. However, the LMS algorithm may never converge to $\mathbf{w}_{opt}$ due to the sample-wise estimation of $E\{ex\}$ by $ex = (\bar{e} + v)x$, where $vx = 0$ does not necessarily hold. 13

5 Model for stereophonic AEC (SAEC). Only two of the four possible echo-paths are shown in the figure not only for simplification purpose but also since the conventional SAEC approach attempts to minimize the mean square error (MSE) for one near-end microphone at a time. A decorrelation procedure is applied to the reference signals before playback and adaptation at the near end to alleviate the non-uniqueness problem. 16

6 Effect of ill-conditioning of the far-end mixing matrix $\mathbf{G}(n)$ on the MSE $E\{|e(n)|^2\}$ for SAEC. The optimal solution is at $\mathbf{h} = [2, 1]^T$ for both (a) and (b). However, even when $\mathbf{G}(n)$ is fully ranked such that there is a unique solution, the deformation of the MSE surface in (b) causes the convergence rate of the LMS-based adaptive algorithm to go down. The worst-case scenario occurs when $\mathbf{G}(n)$ is rank deficient as in (c), for which the solution is a line that goes through $\mathbf{h} = [2, 1]^T$. 18

7 Model for blind source separation (BSS) consisting of the mixing system with Q source signals followed by the de-mixing system with P sensors. 20

8 PCA versus ICA. The original source signal s_1 and s_2 were generated by two independent uniform distributions. [68] 21

9 Model for SAEC based on semi-blind source separation (SBSS). 23

10 GSM AMR speech coding distortion, measured in terms of the signal-to-distortion ratio (SDR), versus input speech magnitude, measured in terms of the signal loss, for various bit-rate (kbps). 27

11 Echo return loss enhancement (ERLE) from the network-based AEC (using the FBLMS algorithm) as a function of the loudspeaker saturation parameter ρ (smaller ρ indicates greater saturation) and the GSM AMR bit-rate. . . . 27

12 ERLE when there is a sampling rate mismatch between the reference signal and the acoustic echo. The correction was made through re-sampling by interpolation in the time domain by using the coefficients re-use block size of B. 29

13 Adaptive filtering with linear or nonlinear distortion on the true, noise-free response $\bar{d}(n)$. A noise-suppressing memoryless nonlinearity $f(\cdot)$ is applied to the observed filter estimation error $e(n)$ in the feedback-loop in order to suppress the effects of the distortion during the adaptation of a linear filter. 31

14 PDFs (scaled by $\log(\cdot)$) and score functions of s, t, and u when $s = t + u$, $t \sim Gaussian(0, 1)$, and $u \sim Laplacian(0, 1)$. For $|x| < 2$, $\phi_s(x)$ exhibits approximately linear scaling by a factor of 0.5 that accounts for relatively equal contributions from t and u for small observed (noisy) magnitude $|s|$. For $|x| > 2$, $\phi_s(x) \to \phi_u(x)$ very rapidly as $|x| \to \infty$ since the kurtosis is higher for u than t, *i.e.*, the probability of u is greater than that of t for large observed magnitude $|s|$. 38

15 Realization of noise-suppressing nonlinearities (61), (62), (65), (66), and (67) obtained through the MMSE and the MAP estimation procedures when the observed adaptive filter estimation error is modeled additively as $e = \bar{e} + v$. 40

16 Realization of ad-hoc noise-suppressing nonlinearities (68) and (69). 41

17 Comparison between AEC and adaptive noise cancellation (ANC). In both cases, the application of ERN to the filer error $e(n)$ allows an LMS-based adaptive filter to produce the estimate of the target source $s(n)$ that is independent on average from the signal $x(n)$ to be canceled. 48

18 From the regularized NLMS algorithms (rNLMS1, rNLMS2) with the ERN ($f_{\text{MMSE}}^{\text{GL}}$). 54

19 From the regularized NLMS algorithm (rNLMS2) with the ERN ($f_{\text{MMSE}}^{\text{GL}}$). Corresponding average tERLE is provided in the legend. 56

20 Network-based AEC with speech coding distortion for various bit-rates (kbps). Quantization noise starts to take over beyond the signal loss of 25 dB. . . . 57

21 From the NLMS algorithms (NLMS, rNLMS2) with the ERN ($f_{\text{MMSE}}^{\text{GG}}$). GSM AMR speech encoding and decoding at 12.2 kbps bit-rate were applied to the acoustic echo. 58

22 From the regularized FBLMS algorithm (RFBLMS) with the ERN ($f_{\text{MMSE}}^{\text{GG}}$, $f_{\text{MMSE}}^{\text{LG}}$, $f_{\text{MMSE}}^{\text{GL}}$). 61

23 From RFBLMS with $f_{\text{MMSE}}^{\text{GL}}$ for various block iterations *iter* and step-size α. The RIR changes suddenly at 10 seconds due to the displacement of the loudspeaker-microphone enclosure. 62

24 From RFBLMS with $f_{\text{MMSE}}^{\text{GL}}$ or $f_{\text{MAP}}^{\text{GL}}$. 63

25 System integration of adaptive filter and REE. 68

26 Signal delay is linearly varied after resampling frame-wise (a) alternatingly across channels and (b,c,d) simultaneously across channels where every other block is resampled after time reversal and reversed back afterward ($N = 2048$, $R = 1.0004$, $f_s = 16$ kHz). 71

27 Inter-channel coherence (averaged over first 5 seconds of speech). From top to bottom: $N = \infty, 4096, 2048, 1024, 512, 256$. 72

28 Inter-channel coherence comparison after decorrelation. 72

29 Distortion on a speech signal by FDR (normalized with respect to TDR) for $R = 1.0004$, $N = 2048$, and $f_s = 16$ kHz. 75

30 Block-iterative adaptation (BIA) of the filter coefficient vector $\mathbf{w}_{i,j}$, where i is the iteration index and j is the block index. 78

31 Near-end loudspeaker, local, and microphone signals for SAEC. Double-arrows indicate individual speech activity. 80

32 True ERLE (averaged over left and right channels) for combinations of *iter*, B, DTD, and EW without decorrelation. 82

33 Improvement in misalignment without decorrelation. 83

34 True residual echo and tERLE without decorrelation. 83

35 True residual echo and tERLE with NLP. 84

36 True residual echo and tERLE with AWGN. 84

37 True residual echo and tERLE with OSD. 84

38 True residual echo and tERLE with RUD or RBI. 85

39 Improvement in misalignment after decorrelation. 85

40 Misalignment averaged over 20 runs and all echo paths. 88

41 Sub-band tERLE decomposition (stereophonic, averaged over all channels). 90

42 Sub-band misalignment decomposition (stereophonic, averaged over all echo paths). 91

43 Variable resampling ratios R_1, R_2, and R_3 and their corresponding coherence plots, which match the coherence from SBR to that of other decorrelation methods. 93

44 FDR with fixed $\Delta R = 0.0028$ and the corresponding variable resampling ratios R_4 and R_5 that match the coherence of FDR in the high frequency band. 93

45 Sub-band tERLE decomposition (single channel). 97

46 Sub-band misalignment decomposition (single channel). 97

47 Model of the near-end and the far-end mixing systems and the semi-blind source separation (SBSS) system. 101

48 Source activities in the worst-case scenario. 123

49 Performance of SBSS with AWGN decorrelation procedure (with de-mixing matrix constraints and 5 iterations). 125

50 Performance of constrained and unconstrained SBSS (with 5 iterations). . . 126

51 Performance of SBSS with different number of iterations (with de-mixing matrix constraint). 127

52 Performance of SBSS with different EBR (with de-mixing matrix constraint and 5 iterations). 129

53 Performance of SBSS with AWGN noise at near-end microphones (with de-mixing matrix constraint and 5 iterations). 130

54 Performance of SBSS with different FFT frame size (with de-mixing matrix constraint and 5 iterations). 131

55 Performance of SBSS with different ICA step-size η (with de-mixing matrix constraint and 5 iterations). 132

56 Performance of SBSS with online and batch-online implementations (with de-mixing matrix constraint and 5 iterations). 133

57 Performance of SBSS with variation in the active source number (with de-mixing matrix constraint and 5 iterations). 134

58 Performance of SBSS for different microphone configurations (with de-mixing matrix constraint and 5 iterations). 136

59 Performance of SBSS with different microphone configurations and without applying the near-end source separation (with de-mixing matrix constraint and 5 iterations). 137

60 Comparison between SBSS and FBLMS for a real-world scenario. 139

SUMMARY

In the real world, an acoustic signal is almost never present in isolation. The ubiquitous "acoustic ambience" is always full of noise of various kinds. These interferences are detrimental in problems that involve linear system identification since the input-output relationship is much more than what is defined by the linear system. In particular, the conventional acoustic echo cancellation (AEC) framework based on the least mean square (LMS) algorithm by itself lacks the ability to handle many secondary signals, *e.g.*, local speech and background noise, that interfere with the adaptive identification and filtering process. There are also nonlinear interferences, such as those introduced by speech codecs used in telecommunication networks, that can pose more difficulties to AEC than linearly additive noises. We need to address the deficiencies of traditional AEC techniques and develop a set of novel techniques to provide the AEC performance for immersive teleconferencing experience that is robust to many possible distortions.

In this dissertation, we build a foundation for what we refer to as the *system approach* to signal enhancement as we focus on the AEC problem. Such a "system" perspective aims for the integration of individual components, or algorithms, into a cohesive unit for the benefit of the system as a whole to cope with real-world enhancement problems. The standard system identification approach by minimizing the mean square error (MSE) of a linear system is sensitive to distortions that greatly affect the quality of the identification result. Therefore, we begin by examining in detail the technique of using a noise-suppressing nonlinearity in the adaptive filter error feedback-loop of the LMS algorithm when there is an interference at the near end, where the source of distortion may be linear or nonlinear. We provide a thorough derivation and analysis of the error recovery nonlinearity (ERN) that "enhances" the filter estimation error prior to the adaptation to transform the corrupted error's distribution into a desired one, or very close to it, in order to assist the linear

adaptation process. We reveal important connections of the residual echo enhancement (REE) technique to other existing AEC and signal enhancement procedures, where the technique is well-founded in the information-theoretic sense and has strong ties to independent component analysis (ICA), which is the basis for blind source separation (BSS) that permits unsupervised adaptation in the presence of multiple interfering signals. Notably, the single-channel AEC problem can be viewed as a special case of semi-blind source separation (SBSS) where one of the source signals is partially known, *i.e.*, the far-end microphone signal that generates the near-end acoustic echo. Indeed, SBSS optimized via ICA leads to the system combination of the LMS algorithm with the ERN that allows continuous and stable adaptation even during double talk.

Next, we extend the system perspective to the decorrelation problem for AEC that, apart from the system identification issues, can retard the adaptive algorithm's convergence speed and degrade the AEC performance. We show that the REE procedure can be applied effectively in a multi-channel AEC (MCAEC) setting to indirectly assist the recovery of lost AEC performance due to inter-channel correlation, known generally as the "non-uniqueness" problem. In addition, we incorporate other techniques to further boost the REE-based MCAEC performance. Specifically, we develop a new technique of decorrelation by resampling (DBR) that directly alleviates the non-uniqueness problem while introducing minimal distortion to signal quality and statistics. We then illustrate the systematic relationship between DBR and block-iterative adaptation (BIA), or batch adaptation in general, utilized by the REE technique. We derive the frequency-domain resampling (FDR) technique as a computationally efficient way to implement DBR with more design flexibility for controlling the trade-off between signal distortion and decorrelation amount. As an advanced extension of FDR, we come up with the sub-band resampling (SBR) technique, which, given the same degree of decorrelation measured in terms of the coherence, has the potential to not only achieve superior audio quality but also provide better overall AEC performance than other decorrelation procedures. We show that the system approach can also be applied to the multi-delay filter (MDF), which suffers from the inter-block correlation problem. Sub-band analysis of the echo return loss enhancement

and the misalignment illustrates that DBR and BIA work together mutually to recover the cancellation performance lost from inter-channel correlation during MCAEC and that BIA enables the MDF to naturally regain the performance lost from inter-block correlation.

Finally, we generalize the MCAEC problem in the SBSS framework and discuss many issues related to the implementation of an SBSS system. After a deep analysis of the structure of the SBSS adaptation, which is a truly multi-channel approach to AEC rather than the conventional single-channel solution to MCAEC, we propose a constrained batch-online implementation that stabilizes the convergence behavior even in the worst case scenario of a single far-end talker along with the non-uniqueness condition on the far-end mixing system. Specifically, we use a matrix constraint to reduce the effect of the non-uniqueness problem. Experimental results show that high echo cancellation can be achieved just as the misalignment remains relatively low without any pre-processing procedure to decorrelate the far-end signals even for the single far-end talker case.

The proposed techniques are developed from a pragmatic standpoint, motivated by real-world problems in acoustic and audio signal processing. From a theoretical point of view, an orthogonality interpretation to the traditional adaptive procedures is a direct consequence of the mean square error optimization criteria that gives rise to the LMS algorithm. Shortcomings of these methods and criteria, although conceptually simple and practical in terms of computational saving, have been much discussed. On the other hand, the results from explicit statistical independence maximization, particularly in multi-channel array signal processing, have been in many cases surprisingly good and robust, as reported in the recent literatures as well as based on our own experience. Generalization of the orthogonality principle to the system level of an AEC problem allows us to relate AEC to source separation that seeks to maximize the independence, hence implicitly the orthogonality, not only between the error signal and the far-end signal, but rather, among all signals involved. The system approach, for which the REE paradigm is just one realization, enables the encompassing of many traditional signal enhancement techniques in analytically consistent yet practically effective manner for solving the enhancement problem in a very noisy and disruptive acoustic mixing environment.

CHAPTER I

INTRODUCTION

1.1 Motivations

Acoustic echo arises during a telecommunication session since there is usually some degree of coupling between a loudspeaker and a microphone at the near end (or the local end), which results in a signal from the far end (or the remote end) being played out and captured locally and sent back to its originating location (see Figure 1). When the round trip delay is substantial ($\sim$ 200 ms or more), the resulting "echo" becomes highly objectionable and may render the teleconferencing ineffective [40, Chapter 1]. The situation is obviously avoided by using a half-duplex transmission mode to eliminate the echo's return path completely (*i.e.*, hard clipping) or a headset to isolate the earpiece from the microphone. But such solutions are not always desirable or even possible. Many everyday situations clearly demand a full-duplex, hands-free communication, in which case the acoustic echo cancellation (AEC) is needed to attenuate the echo sufficiently without disrupting an on-going conversation. In addition, the usage of multiple loudspeakers and microphones to provide the spatial audio awareness requires effective multi-channel acoustic echo cancellation (MCAEC) among other signal enhancement procedures for natural and immersive telecollaboration [136].

As illustrated in Figure 2, AEC is conventionally cast as a system identification problem in which the room impulse response (RIR) (*i.e.*, the echo path) between a loudspeaker and a microphone is estimated by a linear adaptive filter. The least mean square (LMS) algorithm is commonly used for such a task due to its computational simplicity and adaptability. To cancel the echo, the estimated RIR in the form of a finite impulse finite impulse response (FIR) filter is applied to the far-end signal (*i.e.*, the reference signal), the result of which is a close replica of the far-end signal captured at the near-end microphone that can be subtracted from the near-end microphone signal before being transmitted back to

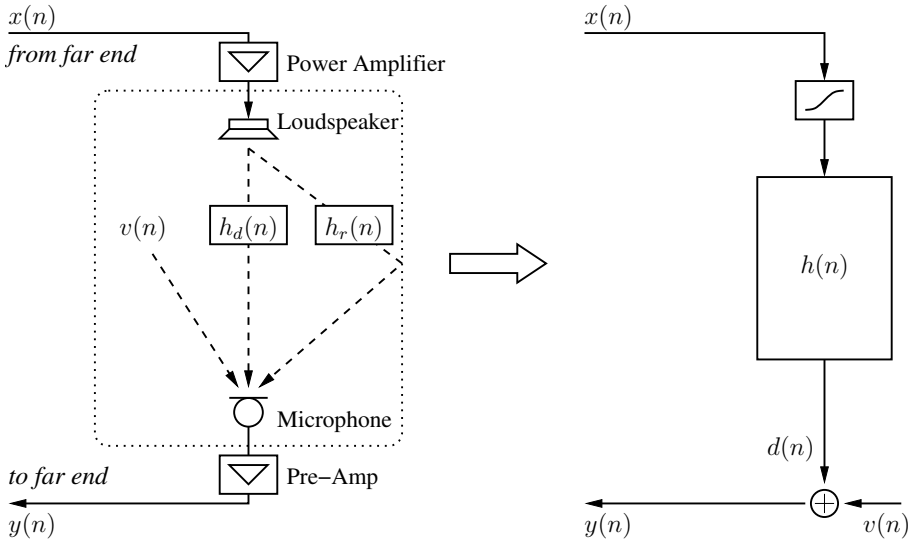


Figure 1: Model for acoustic echo inside the loudspeaker-enclosure-microphone system (LEMS). A loudspeaker power amplifier (PA) is often modeled by a memory-less saturation nonlinearity that distorts the far-end (reference) signal $x(n)$, whereas the effect of a microphone pre-amp is usually ignored. Impulse responses $h_d(n)$, due to a direct echo path, and $h_r(n)$, due to reverberation (*i.e.*, reflections), are combined together as a single room impulse response (RIR) $h(n)$. A local noise $v(n)$ is added directly to the acoustic echo $d(n)$, and the resulting near-end signal $y(n)$ is sent back to the far end.

the far end. However, there may be many types of distortion that can prevent the LMS-based adaptive filter from properly identifying the echo path in the first place for sufficient AEC performance. Any remaining echo then has to be reduced further by residual echo suppression (RES), *e.g.*, a hard-clipping RES.

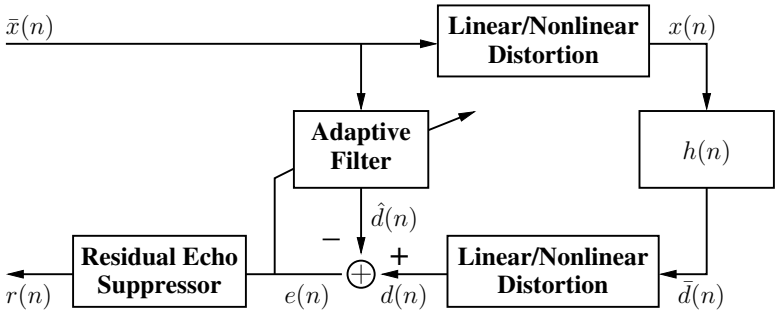


Figure 2: Model for acoustic echo cancellation (AEC). There may be linear or nonlinear distortion applied to the reference signal $\bar{x}(n)$ or the acoustic echo $\bar{d}(n)$ that can prevent a linear finite impulse response (FIR) adaptive filter from properly identifying the RIR $h(n)$. Any remaining echo component in the estimation error $e(n)$ is reduced further through residual echo suppression (RES).

For example, a local noise added to the acoustic echo can complicate the task of system identification since it is unrelated to the reference signal that is actually driving the

loudspeaker-enclosure-microphone system (LEMS) (see Figure 1). Such a type of distortion can be categorized into two groups. One is the ambient background noise, *e.g.*, from an air conditioner or a display projector in a conference room, that is ubiquitous and continuously present. The other type of noise occurs when the near-end talker speaks concurrently with the far-end talker, *i.e.*, the double-talk situation. Although usually short in duration, double talk can severely disrupt the filter adaptation since the near-end speech is colored, non-stationary, and most likely larger in volume than the acoustic echo. A practical yet rather ad-hoc solution, much like the hard-clipping RES approach, is to completely freeze the filter adaptation during double talk, which obviously limits the effectiveness of an adaptive filter to track the changes in the echo path.

Whereas the effect of a local noise is linear in the LEMS, nonlinear characteristics exhibited by a power amplifier (PA), often modeled by a memory-less saturation nonlinearity (see Figure 1), can also degrade the AEC performance excessively [16, 134]. A linear approximation of the LEMS is no longer valid in such a case, and an FIR filter determined by the LMS algorithm is able to cancel only the linear portion of the acoustic echo. In such a case, nonlinear AEC is employed by modeling the LEMS as a Hammerstein system to decouple the echo path into a cascade of nonlinear response, *e.g.*, the loudspeaker saturation, and linear response, *i.e.*, the acoustic impulse response itself (see Figure 1). More detailed treatment of nonlinear AEC can be found in [105].

Another source of nonlinear and pervasive distortion that greatly degrades the AEC performance consists of the speech codecs used in wireless and voice-over-IP (VoIP) communications. Many signal enhancement tasks such as the automatic gain control (AGC) and the noise reduction are often relegated to the telecommunications network providers due to the lack of proper computing resources at the near end, *e.g.*, a cellular handset. Hence the AEC must then be implemented at a central location somewhere in the network as illustrated in Figure 3. When a speech coding distortion is applied to the acoustic echo, even fast-converging alternatives to the LMS algorithm such as the frequency-block LMS (FBLMS) and the recursive least squares (RLS) algorithms are unable to provide an adequate AEC performance by themselves [134].

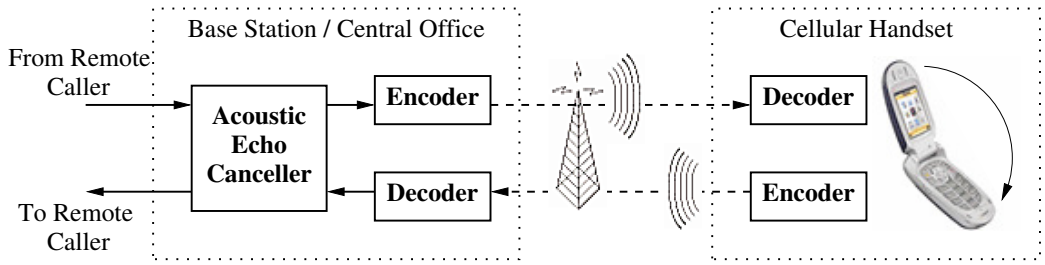


Figure 3: Implementation of AEC in the telecommunications network. Speech coding distortion on the acoustic echo that severely degrades the AEC performance is unavoidable in such a scenario.

Furthermore, a problem that is often overlooked by the manufacturers of audio devices and the developers of signal enhancement algorithms is the mismatch in the sampling rate of digital-to-analog converter (DAC) and analog-to-digital converter (ADC). The sampling rate mismatch of several hundred parts per million (ppm) is normally expected due to the randomness in manufacturing processes and the lack of a common clock source by DAC and ADC, where such a small amount of mismatch can create enough nonlinear distortion to cause a significant degradation in the AEC performance [100].

Finally, a distortion to the acoustic echo may occur simply through a change in the RIR itself, after which an adaptive filter must be able to converge quickly to a new solution. Particularly for the case of MCAEC, a filter needs to be re-adapted also when the far-end echo paths change [118], *e.g.*, when the far-end talkers change in location or when an active talker switches from one to another. The situation is further complicated by the so-called "non-uniqueness" problem caused by the correlated multiple reference signals, *i.e.*, far-end microphone signals, that dramatically slows down the convergence rate of a multi-channel adaptive filter based on the LMS algorithm at the near end.

1.2 Objectives

There are many distortions that can affect the AEC performance as discussed above, including:

- Local noise in the form of ambient background noise.
- Local noise in the form of near-end speech, *i.e.*, the double-talk situation.
- Loudspeaker saturation nonlinearity.

- Speech coding distortion.
- Mismatch in playback and recording sampling rates.
- Ill-conditioning of the far-end mixing system that causes the non-uniqueness problem during MCAEC.

In order to better attack these problems, we may generalize all signal enhancement procedures (*e.g.*, AEC, noise reduction, dereverberation, beamforming, etc.) as "source separation," *i.e.*, estimation of individual signals from a mixture to obtain the signal(s) of interest. AEC and other signal enhancement algorithms are traditionally derived through the standard mean square error (MSE) optimization. This leads to the well-known orthogonality principle, *i.e.*, decorrelation of algorithm output signals. On the other hand, recent works on source separation clearly suggest that by elevating the signal enhancement criterion to achieve maximum statistical independence of the processed signals, an improved and even more robust performance can be obtained than with standard MSE-based methods. By "robustness," we mean the ability to provide sufficient performance even if the assumed conditions deviate from the ideal ones. The conventional MCAEC framework happens to be inherently limited by the single-channel LMS optimization approach that does not properly address the robustness issue.

The main challenge is in deriving a new AEC paradigm in the framework of blind source separation (BSS) based on independent component analysis (ICA) that provides a solid analytical foundation to solving the robustness issue. Many classical AEC algorithms must be adjusted often in an ad-hoc fashion to fit the real-world situations and, by lacking the robustness, may consequently fail to produce the desired results. The ICA-based BSS is a very powerful and effective signal enhancement method that allows the recovery of a target signal from a mixture of multiple signals even when the original signals are unavailable (hence "blind"). By relating the LMS-based AEC to the ICA-based semi-blind source separation (SBSS), in which case some of the source signals are known in advance (*e.g.*, the far-end reference signal), the development of many new AEC techniques that are much superior to the pre-exiting ones should be realized.

Our overall objective is to establish what we call the *system approach* for the development of algorithms, or components, for robust AEC in order to overcome the real-world factors. A "system" is defined here to be an integration of individual components that are designed properly to interact with each other mutually for the benefit of the system as a whole. Specifically, we focus on the following main topics in this dissertation.

1. Robust AEC through residual echo enhancement (REE).

2. System perspective of decorrelation for AEC.

3. Generalization of MCAEC by the SBSS system.

The REE technique is just one realization of the system approach to the AEC problem that, in fact, falls out naturally from the SBSS framework. The goal is to show that by raising the orthogonality principle to the system level that seeks to maximize the independence between outputs representing all the involved signals, robust signal enhancement is possible in a very noisy and disruptive acoustic mixing environment.

1.3 Outline

The rest of this dissertation is organized as follows.

In Chapter 2, we briefly go over adaptive filtering algorithms based on the MSE optimization for single-channel AEC. We review the LMS algorithm and the corresponding orthogonality principle in detail. We discuss MCAEC and the non-uniqueness problem, which occurs when the single-channel LMS algorithm is applied to the MCAEC problem. We then provide an overview of BSS, the statistical independence maximization via ICA, and SBSS for a generalized framework for MCAEC. We also define the performance measures for AEC and the coherence, or correlation, measure for MCAEC, and examine the effect of linear and nonlinear distortions on the AEC performances.

In Chapter 3, we establish the foundation for the system approach to robust AEC by introducing the REE technique. We derive what we refer to as the error recovery nonlinearity (ERN) through the Bayesian estimation procedure for the suppression of additive noise remaining in an adaptive filter's cancellation error that allows the adaptive filter to cope

with distortions to the acoustic echo. We show that the technique is fundamentally related to many other conventional procedures for the robustification of AEC, and that it is also well-founded in an information-theoretic sense with deep connections to ICA and SBSS. Through the system approach, we motivate what we call the block-iterative adaptation (BIA) as an essential part of the robust AEC system that not only permits the recovery of the convergence speed lost in noisy situations but also the reduction in sensitivity to the mis-estimation of signal statistics. We implement the REE procedure on single-channel AEC simulation with linear and nonlinear distortions to illustrate the technique's effectiveness.

In Chapter 4, we expand the system perspective to cover the decorrelation aspect for assisting the AEC process. We first apply the REE technique to MCAEC to show that BIA permits a natural recovery of the AEC performance lost due to inter-channel correlation during MCAEC. We propose the novel decorrelation-by-resampling (DBR) technique that is compatible with the REE-based MCAEC system for alleviating the non-uniqueness problem. We then extend the DBR technique to frequency-domain resampling (FDR) for computational saving, where we obtain the sub-band resampling (SBR) technique as a variation of FDR that not only achieves superior audio quality compared to other decorrelation procedures but also provides the most improvement in the overall AEC performance. Finally, we apply the decorrelation via system approach to the multi-delay filter (MDF) [120], which suffers from inter-block correlation [19]. We demonstrate the advantage of the system approach through the sub-band decomposition of the true echo return loss enhancement (tERLE) and the misalignment obtained from MCAEC and MDF simulations.

In Chapter 5, we make a detailed investigation into SBSS as the generalization of MCAEC. We first define the SBSS model and cast the non-uniqueness problem in the SBSS framework. We then derive the steady-state state solution for SBSS and make a clear connection between the MSE solution to MCAEC and the ICA solution to SBSS, where we show that the ICA solution is contingent on certain constraints on the separation matrix. As the convergence rate of ICA-based adaptive algorithms is generally much slower compared to that of the LMS algorithm and is strongly dependent on the amount of data, we introduce the batch-online adaptation procedure for practical SBSS and provide an outline

of the proposed SBSS algorithm along with simulation results.

Finally in Chapter 6, we conclude by presenting the overall summary of our research, a list of specific contributions, and a list of future research topics.

CHAPTER II

PROBLEM BACKGROUND

This chapter is organized as follows. First in Section 2.1, we go over the acoustic echo cancellation (AEC) problem and its workhorse, the least mean square (LMS) algorithm based on the mean square error (MSE) optimization. Second in Section 2.2, we present the multi-channel AEC (MCAEC) scenario and the corresponding non-uniqueness problem that occurs when the LMS algorithm is applied to MCAEC. Third in Section 2.3, we motivate blind source separation (BSS), or namely semi-blind BSS (SBSS), for the generalization of MCAEC that maximizes the statistical independence of output signals via independent component analysis (ICA). Finally in Section 2.4, we examine the effect of linear and nonlinear distortions on linear adaptive filtering algorithms.

2.1 Acoustic Echo Cancellation (AEC)

2.1.1 Mean Square Error (MSE) Minimization

A linear adaptive filter can be obtained through an optimization procedure that minimizes some cost function (or criterion) in terms of the estimation error and the filter coefficients. There are basically three types of sample-based time-domain adaptive filtering algorithms commonly used for AEC that are derived from the traditional least-square optimization approach: LMS [138, 139], affine projection (AP) [94, 40], and recursive least squares (RLS) [49, 46]. The corresponding cost function and algorithmic complexity for the three algorithms are listed in Table 1.

The convergence rate of the AP and the RLS algorithms is generally much higher than that of the LMS algorithm, but so is their computational complexity. The LMS algorithm is a special case of the AP algorithm when only a single tap rather than multiple taps of past estimation error is used to update the filter coefficients. The LMS algorithm is also a robust way to stochastically approximate the least-squares solution [47] and is derived in the same fashion as with the Wiener filter through the minimization of the MSE $E\{|e(n)|^2\}$, where

Table 1: Three sample-based adaptive algorithms obtained through least-square optimization and commonly used for AEC: least mean square (LMS), affine projection (AP), and recursive least squares (RLS). $E\{\cdot\}$ is the expectation operator, $e(n)$ is the estimation error, L is the filter order, P is the AP order, and $0 < \lambda \leq 1$ is the forgetting factor.

Algorithm	Cost Function	Complexity
LMS	$E\{\|e(n)\|^2\}$	$O(L)$
AP	$E\{\sum_{i=0}^{P-1} \|e(n-i)\|^2\}$	$O(LP)$
RLS	$\sum_{i=0}^{n} \lambda^{n-i}\|e(i)\|^2$	$O(L^2)$

$E\{\cdot\}$ is the expectation operator and $e(n)$ is the filter output error. The RLS algorithm, which may be considered as a stochastic version of the Kalman filter and exhibits much faster convergence speed than the Wiener filter, is computationally far more expensive than the LMS algorithm and is susceptible to numerical instability since it recursively estimates the inverse of the input auto-correlation matrix. The LMS algorithm suffers the most from the decreased convergence rate when the input signal is non-white, *e.g.*, the speech signal. Still, the LMS algorithm is widely adopted for the AEC purpose due to its computational efficiency and adaptability, and it is a standard against which all other algorithms' performances are measured.

On the other hand, block-based frequency-domain versions of the LMS algorithm include frequency-block LMS (FBLMS) [23, 115], multi-delay filter [120] (or equivalently, partitioned block frequency-domain adaptive filter (PBFDAF) [17]), and generalized MDF (GMDF) [82]. These adaptive algorithms achieve improved computational efficiency and faster convergence rate than the time-domain sample-based counterpart through the implementation of filtering (*i.e.*, convolution) and adaptation in the frequency domain. Although the overlap-add (OLA) or overlap-save (OLS) method [93] can be used to reduce a buffering delay introduced by the block processing, the delay may still be significant when a long room impulse response (RIR) corresponding to the reverberant acoustic condition must be estimated. A trade-off between the computational complexity, convergence rate, and processing delay must be considered while implementing the frequency-domain AEC in real time.

2.1.2 Least Mean Square (LMS) Algorithm

The LMS algorithm is given by the adaptation rule

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu e(n)\mathbf{x}(n), \tag{1}$$

where $\mathbf{x}(n) = [x(n), x(n-1), \cdots, x(n-L+1)]^T$ is the reference signal vector of length L at time n, $\mathbf{w}(n) = [w_0(n), w_1(n), \cdots, w_{L-1}(n)]^T$ is the filter coefficient vector of the same length, "^{T}" is the transposition operator, μ is the adaptation step-size parameter, and

$$e(n) = d(n) - \hat{d}(n) \tag{2}$$

is the error from approximating the observed acoustic echo $d(n)$ by the replica

$$\hat{d}(n) = \mathbf{w}^T(n)\mathbf{x}(n) \tag{3}$$

based on the estimated echo return path response reflected in $\mathbf{w}(n)$. It iteratively and stochastically solves the normal, or Wiener-Hopf, equation

$$\mathbf{R}_{xx}\mathbf{w} = \mathbf{r}_{xd}, \tag{4}$$

where $\mathbf{R}_{xx} = E\{\mathbf{x}\mathbf{x}^T\}$ is the auto-correlation matrix and $\mathbf{r}_{xd} = E\{\mathbf{x}d\}$ is the cross-correlation vector.

There are many real-world factors that affect the performance of the LMS algorithm. First and the foremost is that a finite number of coefficients are used to describe a system response $h(n)$ that is usually infinite in length, hence $e(n) \neq 0$ no matter how perfect an adaptive algorithm is. Another factor is that the approximation of the gradient of the MSE by a sample value, $\nabla_{\mathbf{w}} E\{e(n)^2\} = -2E\{e(n)\mathbf{x}(n)\} \approx -2e(n)\mathbf{x}(n)$, leads to a "noisy" update (*i.e.*, the gradient noise [49]) of the filter coefficients, which exacerbates the sensitivity of the adaptive algorithm to non-whiteness and non-stationarity of signals. Finally, there are often local noises that corrupt the acoustic echo, *i.e.*,

$$d(n) = \bar{d}(n) + v(n), \tag{5}$$

$$\bar{d}(n) = h(n) * x(n) \approx \mathbf{h}^T(n)\mathbf{x}(n), \tag{6}$$

where $\bar{d}(n)$ is the *true* acoustic echo, $v(n)$ is the noise, assumed here to be additive at the time-sample level, "$*$" is the convolution operator, and $\mathbf{h}(n)$ is a truncated vector representation of the RIR $h(n)$. By "true," we mean a distortion-free situation, *i.e.*, $v(n) = 0$. The resulting estimation error then becomes

$$e(n) = \bar{e}(n) + v(n), \tag{7}$$

$$\bar{e}(n) = \bar{d}(n) - \hat{d}(n) \approx \mathbf{h}_{\Delta}^{T}(n)\mathbf{x}(n), \tag{8}$$

where $\bar{e}(n)$ is the true error, or residual echo, and

$$\mathbf{h}_{\Delta}(n) = \mathbf{h}(n) - \mathbf{w}(n) \tag{9}$$

is the misalignment (or system distance [46]) vector [1].

After its introduction in [138], the LMS algorithm was initially applied to the network echo (or line echo) cancellation (NEC) in the 1960s [119, 116]. A network echo occurs due to the reflection of electrical signals caused by the impedance mismatch between the two-wire and the four-wire circuits in the telecommunications network. AEC is much more difficult to implement than NEC since the acoustic echo is usually longer in length (on the order of 64 to 128 ms or more with the reverberation time of $T_{60} > 200$ ms) and highly time-varying when compared to the network echo (typically less than 64 ms) [40, 9]. Many modifications have been proposed to make the LMS algorithm robust to deviations from ideal conditions necessary for optimal system identification during AEC.

2.1.3 Orthogonality Principle

A major deficiency of the LMS algorithm is that the short-term effect of additive noise on the algorithm is overlooked from the sample-based optimization process.

To begin with, we observe that the LMS algorithm of (1) is actually a stochastic version of the LMS gradient descent algorithm

$$\mathbf{w}(n+1) = \mathbf{w}(n) - \mu \nabla_{\mathbf{w}} E\{|e(n)|^2\}, \tag{10}$$

[1] For simplification purpose, the RIR $h(n)$ is assumed to be finite in length, represented by a vector $\mathbf{h}(n)$ of length N such that $\bar{d}(n) = \mathbf{h}^T(n)\mathbf{x}(n)$, and also $N = L$ is assumed if not otherwise specified. To simplify the statistical analysis, all signals are assumed to be generated by zero-mean random processes.

where the direction and the amount of adaptive correction to $\mathbf{w}(n)$ depend on the sample estimate of the gradient

$$\nabla_{\mathbf{w}} E\{|e(n)|^2\} = -2E\{e(n)\mathbf{x}(n)\}. \tag{11}$$

(10) converges to the optimal filter coefficients $\mathbf{w}_{opt}(n)$ when $E\{e(n)\mathbf{x}(n)\} = \mathbf{0}$ that minimizes the MSE and consequently satisfies the orthogonality principle $E\{e(n)x(n)\} = 0$ [49]. The orthogonality principle is illustrated in Figure 4, where a projection of the desired signal $\bar{d}(n)$, which resides in the L-dimensional reference signal space, onto the reference signal vector $\mathbf{x}(n)$ gives the optimal estimate $\hat{d}_{opt}(n)$ such that the true estimation error $\bar{e}(n)$ is orthogonal, *i.e.*, uncorrelated, to $x(n)$. The error is modeled to be quadratic in the space of the parameter $\mathbf{w}(n)$, and the adaptation rule is based on the second-order statistics (SOS).

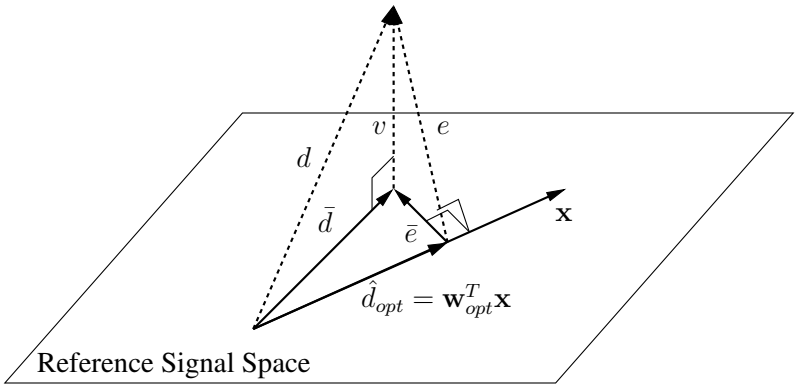


Figure 4: Effect of additive noise v on the least mean square (LMS) estimation of the optimal filter coefficient vector $\mathbf{w}_{opt}$. The optimal solution is obtained as long as $E\{vx\} = 0$, *i.e.*, $E\{e\} = E\{(\bar{e}+v)\mathbf{x}\} = E\{\bar{e}x\}$. However, the LMS algorithm may never converge to $\mathbf{w}_{opt}$ due to the sample-wise estimation of $E\{ex\}$ by $ex = (\bar{e}+v)x$, where $vx = 0$ does not necessarily hold.

As Figure 4 also shows, (11) implies that the optimal solution is attainable in principle even when $v(n) \neq 0$ as long as $v(n)$ is uncorrelated with $x(n)$ such that $E\{v(n)\mathbf{x}(n)\} = \mathbf{0}$ and

$$\begin{aligned}\nabla_{\mathbf{w}} E\{|e(n)|^2\} &= -2E\{(\bar{e}(n) + v(n))\mathbf{x}(n)\} \\ &= -2E\{\bar{e}(n)\mathbf{x}(n)\}.\end{aligned} \tag{12}$$

However, the orthogonality condition is undermined when the expectation in (11) is estimated by a sample value in the LMS algorithm, *i.e.*,

$$\begin{aligned}\nabla_{\mathbf{w}} E\{|e(n)|^2\} &\approx \nabla_{\mathbf{w}} |e(n)|^2 \\ &= -2e(n)\mathbf{x}(n) \\ &= -2(\bar{e}(n)\mathbf{x}(n) + v(n)\mathbf{x}(n)).\end{aligned} \tag{13}$$

That is, even though the noise is most likely uncorrelated with the reference signal, in which case the optimal solution should be still attainable since $E\{e(n)\mathbf{x}(n)\} = E\{\bar{e}(n)\mathbf{x}(n)\} + E\{v(n)\mathbf{x}(n)\} = E\{\bar{e}(n)\mathbf{x}(n)\}$ (*i.e.*, $v(n)$ and $x(n)$ are uncorrelated on average), such an ideal condition does not always hold for individual samples that may actually be correlated in a short period of time such that $v(n)\mathbf{x}(n) \neq \mathbf{0}$, which translates to noisy updating of $\mathbf{w}(n)$ at each iteration. This causes the LMS algorithm, which has an inherent ability to eventually converge, with an arbitrary degree of precision depending on the filter length, to a unique and optimal solution $\mathbf{w}_{opt}(n)$, to have difficulty in converging to the optimal solution in a noisy situation. Therefore, coupled with the gradient noise, the local noise can greatly disrupt the LMS adaptation process and contribute to even larger MSE.

2.1.4 Performance Measures

The two conventional performance measures for AEC are the echo return loss enhancement (ERLE),

$$\text{ERLE (dB)} \equiv 10\log_{10}\frac{E\{|d(n)|^2\}}{E\{|e(n)|^2\}} \approx 10\log_{10}\frac{\sum_n |d(n)|^2}{\sum_n |e(n)|^2}, \tag{14}$$

i.e., the ratio of energies between the acoustic echo $d(n)$ and the residual echo $e(n)$, and the normalized misalignment (or simply "misalignment"),

$$\text{Misalignment (dB)} \equiv 10\log_{10}\frac{E\{|h_\Delta(n)|^2\}}{E\{|h(n)|^2\}} \approx 10\log_{10}\frac{\|\mathbf{h}_\Delta(n)\|^2}{\|\mathbf{h}(n)\|^2}, \tag{15}$$

i.e., the ratio of energies between the misadjustment $h_\Delta(n)$ and the RIR $h(n)$. The ERLE, which should be as high as possible, measures the MSE performance $E\{|e(n)|^2\}$, while the misalignment, which should be as low as possible, measures the mean square deviation (MSD) performance $E\{|h_\Delta(n)|^2\}$.

When $v(n) \neq 0$, $e(n) \rightarrow v(n)$ as $n \rightarrow \infty$ by (7), and the ERLE defined in (14) converges to the *a posteriori* SNR,

$$\textit{a posteriori}\ \text{SNR (dB)} \equiv 10\log_{10}\frac{E\{|d(n)|^2\}}{E\{|v(n)|^2\}} \approx 10\log_{10}\frac{\sum_n |\bar{d}(n)+v(n)|^2}{\sum_n |v(n)|^2}, \tag{16}$$

which is larger than the *a priori* SNR,

$$\textit{a priori}\ \text{SNR(\ dB)} \equiv 10\log_{10}\frac{E\{|\bar{d}(n)|^2\}}{E\{|v(n)|^2\}} \approx 10\log_{10}\frac{\sum_n |\bar{d}(n)|^2}{\sum_n |v(n)|^2}. \tag{17}$$

Hence the standard definition of the ERLE tends to hide the true echo cancellation performance when there is an additive noise. A more objective MSE performance measure is what we refer to as the true ERLE (tERLE),

$$\text{tERLE (dB)} \equiv 10\log_{10}\frac{E\{|d(n)-v(n)|^2\}}{E\{|e(n)-v(n)|^2\}} \approx 10\log_{10}\frac{\sum_n |\bar{d}(n)|^2}{\sum_n |\bar{e}(n)|^2}, \tag{18}$$

i.e., the usual ERLE measured after subtracting out the contributions from the near-end noise.

2.2 Multi-channel AEC (MCAEC)

2.2.1 Single-channel Solution to MCAEC

Figure 5 illustrates the case of a simplest stereophonic AEC (SAEC) (*i.e.*, two loudspeakers and two microphones) with one talker located at the far end and another located at the near end. Any adaptive algorithm may be used to minimize the MSE individually for each near-end microphone channel's error $e(n) = d(n) - \hat{d}(n) = (\mathbf{h}(n) - \mathbf{w}(n))^T\mathbf{x}(n)$, where $\mathbf{h}(n) = [\mathbf{h}_1^T(n), \mathbf{h}_2^T(n)]^T$, $\mathbf{w}(n) = [\mathbf{w}_1^T(n), \mathbf{w}_2^T(n)]^T$, and $\mathbf{x}(n) = [\mathbf{x}_1^T(n), \mathbf{x}_2^T(n)]^T$ are formed by concatenating the respective vectors from the two channels. Thus MCAEC is traditionally approached as a single-channel problem, which consequently leads to the same noise-robustness issues associated with the single-channel AEC.

Furthermore, the reference signals $x_1(n)$ and $x_2(n)$ are highly correlated when they are produced by a single far-end source. In such a case, the auto-correlation matrix $\mathbf{R}_{xx}(n) = E\{\mathbf{x}(n)\mathbf{x}^T(n)\}$ formed by $\mathbf{x}(n)$ is very poorly conditioned, which in effect slows down the convergence rate of an LMS-based adaptive filter. The conditioning of $\mathbf{R}_{xx}(n)$ is improved naturally when there are two or more active far-end sources, but there is normally at most

one active talker on one end during a conversation. The worst-case scenario of ranked-deficient $\mathbf{R}_{xx}(n)$ occurs when the adaptive filter length L is longer than or equal to the far-end RIR length M that leads to the non-uniqueness problem, in which case no unique solution exists for the echo paths $h_1(n)$ and $h_2(n)$ [118].

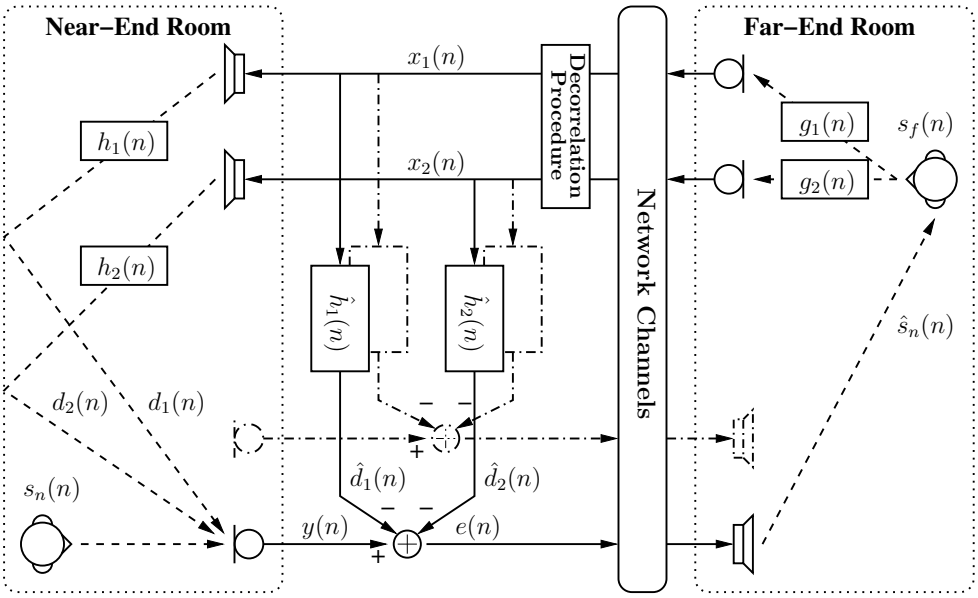


Figure 5: Model for stereophonic AEC (SAEC). Only two of the four possible echo-paths are shown in the figure not only for simplification purpose but also since the conventional SAEC approach attempts to minimize the mean square error (MSE) for one near-end microphone at a time. A decorrelation procedure is applied to the reference signals before playback and adaptation at the near end to alleviate the non-uniqueness problem.

Since the acoustic impulse response in reality is infinite in length, the uniqueness condition of $L < M$ should automatically be satisfied. Still, severe ill-conditioning of $\mathbf{R}_{xx}(n)$ occurs when there is only one colored source at the far end. A high level of inter-channel correlation and the effect of local distortions would then cause the convergence rate of an LMS-based adaptive algorithm to decrease so much that the solution would appear as if it is non-unique. Also, the tail effect occurs when the RIR is under-modeled in length (*i.e.*, $L < N$) such that it leads to the biased filter coefficients when $\mathbf{R}_{xx}(n)$ is ill-conditioned even if a unique solution exists [9, Chapter 5].

2.2.2 Non-uniqueness Problem

The origin of the non-uniqueness problem in MCAEC is generalized as follows.

As shown in Figure 5 for the case of SAEC, let $s_f(n)$ be the far-end source signal. For $i = 1, 2, \ldots, P$, where P is the number of channels, let $\mathbf{g}_i(n)$ be the i^{th} far-end RIR vector of length M, $\mathbf{s}_f(n)$ be the far-end source signal vector of same length, $x_i(n) = \mathbf{g}_i^T(n)\mathbf{s}_f(n)$ be the i^{th} reference signal, and $\mathbf{h}_i(n)$ be the i^{th} near-end RIR vector of length $N = L$ (*i.e.*, same length as $\mathbf{w}_i(n)$). Also, let $s_n(n) = 0$ for the near-end source signal (*i.e.*, no double talk). Then we can show (under regular assumptions) that the MSE is given by

$$\begin{aligned} E\{|e(n)|^2\} &= \mathbf{h}_\Delta^T(n)\mathbf{R}_{xx}(n)\mathbf{h}_\Delta(n) \\ &= \mathbf{h}_\Delta^T(n)\mathbf{G}(n)\mathbf{R}_{s_f s_f}(n)\mathbf{G}^T(n)\mathbf{h}_\Delta(n), \end{aligned} \tag{19}$$

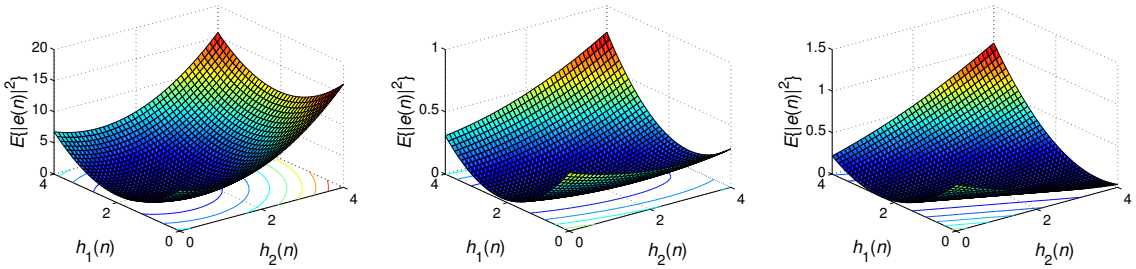
where $\mathbf{h}(n) = [\mathbf{h}_1^T(n), \ldots, \mathbf{h}_P^T(n)]^T$, $\mathbf{w}(n) = [\mathbf{w}_1^T(n), \ldots, \mathbf{w}_P^T(n)]^T$, $\mathbf{h}_\Delta(n) = \mathbf{h}(n) - \mathbf{w}(n)$, $\mathbf{x}(n) = [\mathbf{x}_1^T(n), \ldots, \mathbf{x}_P^T(n)]^T$, $\mathbf{R}_{xx} = E\{\mathbf{x}(n)\mathbf{x}^T(n)\}$, $\mathbf{R}_{s_f s_f} = E\{\mathbf{s}_f(n)\mathbf{s}_f^T(n)\}$, and $\mathbf{G}(n)$ is a $PL \times (L + M - 1)$ matrix with rank PL that is formed by delayed versions of $\mathbf{g}_i(n)$ (refer to [57] for a more detailed presentation). Assuming that $\mathbf{R}_{s_f s_f}$ is fully ranked, the optimal solution for $\mathbf{h}(n)$ is given by

$$\mathbf{h}_\Delta^T(n)\mathbf{G}(n) = (\mathbf{h}(n) - \mathbf{w}(n))^T\mathbf{G}(n) = \mathbf{0}, \tag{20}$$

which is a system of PL unknown coefficients and $(P-1)L + M - 1$ equations. (20) is an exact or at least an over-determined problem as long as $PL \leq (P-1)L + M - 1$ and has a unique solution if and only if $L < M$. Therefore, the non-uniqueness problem arises when the matrix $\mathbf{G}(n)$ is either ill-conditioned or rank deficient. (20) also shows clearly that a change in the far-end mixing matrix $\mathbf{G}(n)$ (*e.g.*, when the far-end talker activity switches) would cause an adaptive algorithm to re-converge to a new solution.

Figure 6 is a simplified illustration of the effect of the far-end mixing system on the MSE for SAEC when $L = N = 1$, $M = 2$, and $\mathbf{h}(n) = [2, 1]^T$. The MSE surface becomes deformed by the ill-conditioning of the far-end mixing matrix $\mathbf{G}(n)$, which in effect decreases the convergence rate of the LMS-based adaptive algorithm (see Figure 6(b)). When $\mathbf{G}(n)$ is rank deficient, the dimension of the solution space of $\mathbf{h}(n)$ (or alternatively the null space of $\mathbf{R}_{xx}(n)$) is greater than zero such that there is no unique solution (see Figure 6(c)). For the case of $L = M = N \geq 2$, the MSE surface forms a "manifold" and cannot be

visualized easily in a three-dimensional space [32]. Figure 6 can be used to visualize the non-uniqueness problem in the frequency domain, in which the representation of a mixing system by a linear model is valid at each frequency bin for any P.



(a) $\mathbf{G}(n) = \mathbf{I}$. (b) $\mathbf{G}(n) \neq \mathbf{I}$ but fully ranked. (c) $\mathbf{G}(n)$ is rank deficient.

Figure 6: Effect of ill-conditioning of the far-end mixing matrix $\mathbf{G}(n)$ on the MSE $E\{|e(n)|^2\}$ for SAEC. The optimal solution is at $\mathbf{h} = [2, 1]^T$ for both (a) and (b). However, even when $\mathbf{G}(n)$ is fully ranked such that there is a unique solution, the deformation of the MSE surface in (b) causes the convergence rate of the LMS-based adaptive algorithm to go down. The worst-case scenario occurs when $\mathbf{G}(n)$ is rank deficient as in (c), for which the solution is a line that goes through $\mathbf{h} = [2, 1]^T$.

Many decorrelation procedures have been proposed to reduce the inter-channel correlation during SAEC (see [9, 40] for references), one of which applies memory-less nonlinearities to the reference signals before playback and adaptation at the near end [12]. Addition of random noises to the reference signals has also proven to be effective. However, these techniques come with the cost of degraded signal quality and must be controlled carefully to not destroy the spacial audio information perceived by the near-end listeners. Other indirect ways for reducing the inter-channel correlation include the input-sliding technique [125], the selective-tap adaptive algorithm [64], and the two-stage estimation of the echo path [31], where all these methods introduce some form of distortion to the estimated echo paths to speed up the convergence rate of MSE-based adaptive algorithms.

2.2.3 Coherence Measure

A standard measure for quantifying the amount of correlation between two signals $x_1(n)$ and $x_2(n)$ is the magnitude-squared coherence (MSC) [20] (or "coherence" in short),

$$C_{x_1x_2}(k) = \frac{|E\{X_1(k)X_2^*(k)\}|}{|E\{X_1(k)\}||E\{X_2(k)\}|}, \tag{21}$$

where $X_1(k)$ and $X_2(k)$ are the discrete Fourier transform (DFT) of $x_1(n)$ and $x_2(n)$, respectively, and "*" is the complex conjugate operator. It can be shown for SAEC that the misalignment at each frequency bin k is inversely proportional to $1-C_{x_1x_2}(k)$ [36, 63, 62].

2.3 Source Separation

2.3.1 Blind Source Separation (BSS)

BSS refers to a signal enhancement procedure that attempts to recover the original source signals when only the observations of some mixture of the signals are available. The simplest scenario is when the mixing system is linear and instantaneous as indicated in Figure 7. That is, given the mixing matrix $\mathbf{A}(n)$ and the source signals vector $\mathbf{s}(n)$, the observed signals are given by the vector

$$\mathbf{x}(n) = \mathbf{A}(n)\mathbf{s}(n). \tag{22}$$

The problem is to obtain the de-mixing matrix $\mathbf{W}(n)$ such that

$$\mathbf{d}(n) = \mathbf{W}(n)\mathbf{x}(n) \simeq \mathbf{s}(n). \tag{23}$$

The matrix $\mathbf{W}(n)$ can be found up to arbitrary permutation and scaling of rows as long as the original sources are statistically independent from one another and $\mathbf{A}(n)$ is an invertible square matrix [56]. In general, a unique solution still exists if the number of sources Q is less than the number of sensors P, *i.e.*, the over-determined case. The under-determined case of $Q > P$ is a difficult problem and is out of the scope of this research. Without loss of generality, $Q = P$ and the invertibility of $\mathbf{A}(n)$ are assumed hereafter.

However, the acoustic mixing systems in real life are convolutive by nature and not instantaneous. To solve such an obstacle, the separation is performed in the DFT domain, as a convolutive mixture in the time domain becomes an instantaneous mixture in the frequency domain:

$$\mathbf{y}(k,l) = \mathbf{A}(k)\mathbf{s}(k,l), \tag{24}$$

where k is the frequency index and l is the block index in time. A major drawback is that the permutation ambiguity exists at each frequency, and the separated components must somehow be re-aligned and associated to correct sources for all frequencies. There are

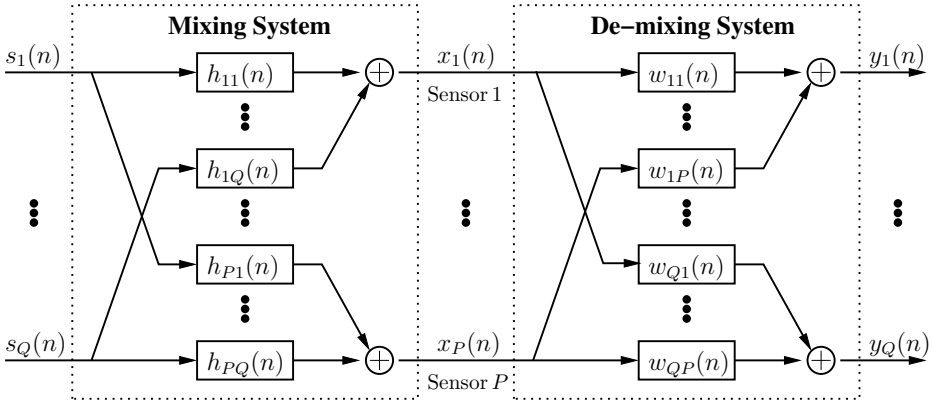


Figure 7: Model for blind source separation (BSS) consisting of the mixing system with Q source signals followed by the de-mixing system with P sensors.

many on-going researches on the permutation alignment problem in the frequency domain. One set of approaches consists of methods based on the time difference of arrival (TDOA) or direction of arrivals (DOA) [102, 88]. Another set consists of those based on the inter-frequency magnitude correlation and the spectral continuity across frequency, both of which applies directly to the speech signal [98, 87]. The scaling ambiguity is much simpler to solve than the permutation ambiguity and can be corrected through either the projection-back processing [86] or the minimum distortion principle (MDP) [78]. Good coverage of the ambiguity problems and other techniques associated with BSS are provided in [73].

2.3.2 Independence Maximization via Independent Component Analysis (ICA)

A very powerful and effective approach for estimating the de-mixing matrix $\mathbf{W}(n)$ when sources are non-Gaussian is ICA that utilizes the higher-order statistics (HOS) [24]. The object is to maximize the statistical independence between the separated output signals given that the source signals are independent to start with.

A standard adaptive algorithm used for ICA optimization is the natural gradient (NG) algorithm, which is expressed in the frequency domain as

$$\mathbf{W}_{m+1}(k) = \mathbf{W}_m(k) + \mu \left(\mathbf{I} - E\{\phi(\mathbf{y}(k,l))\mathbf{y}^T(k,l)\}\right) \mathbf{W}_m(k), \tag{25}$$

where m is the iteration index, μ is the step-size, $\mathbf{I}$ is the identity matrix, and $\phi(\cdot)$ is a memory-less nonlinearity commonly referred to as the score function that depends on the

source's probability density function (PDF). (25) converges when $E\{\phi(\mathbf{y}(k,l))\mathbf{y}^T(k,l)\} = \mathbf{I}$, *i.e.*, the off-diagonal terms of the nonlinear cross-correlation matrix $E\{\phi(\mathbf{y}(k,l))\mathbf{y}^T(k,l)\}$ become zero. The process is similar to principal component analysis (PCA) that uses the SOS to obtain decorrelated components, and (25) can be interpreted as performing the nonlinear PCA to decorrelate the mixed signals [56].

The advantage of ICA over PCA is illustrated in Figure 8 [68]. By maximizing the variance in the direction of the principal components, thereby minimizing the correlation between the unmixed signals, PCA is able to recover the best estimate of the original signals. However, the scatter plots show that while PCA is able to "sphere" to decorrelate, it is unable to "rotate" to match the directions of the latent components. This is since the uniform distribution requires not just the SOS but also the HOS for its complete characterization. On the other hand, ICA is capable of achieving much better separation than PCA for non-Gaussian signals (*e.g.*, speech) by taking into account all the HOS. The disadvantage is that then at most one of the source signals can be Gaussian distributed for ICA to be effective [56].

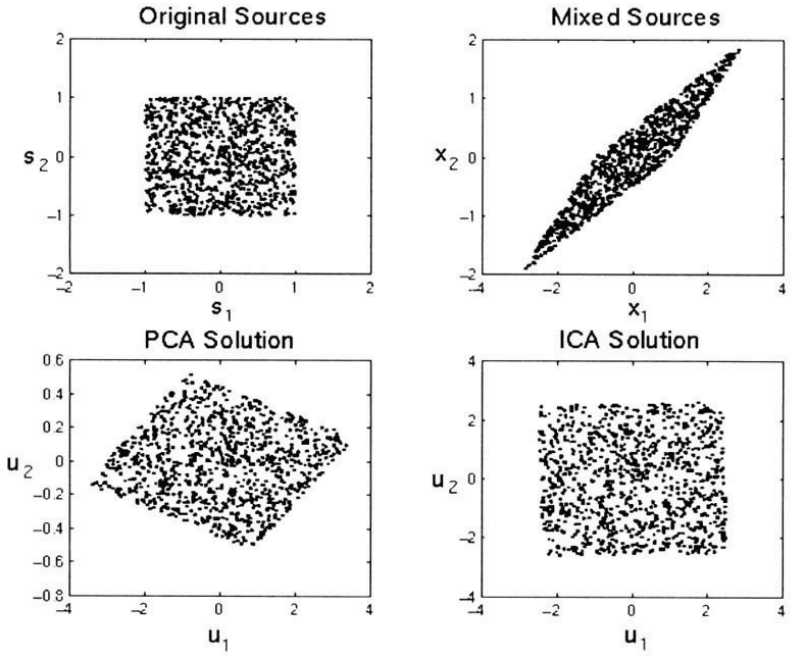


Figure 8: PCA versus ICA. The original source signal s_1 and s_2 were generated by two independent uniform distributions. [68]

There are several ways to derive the NG algorithm (see [68, 56] for references). One approach is to maximize the transfer of statistical information during the separation process, *i.e.*, the infomax or equivalently the maximum-likelihood (ML) approach [6]. Another way is to minimize the mutual information (MI) or equivalently to maximize the statistical independence between the separated components [68]. Differences in the geometrical interpretation between the ML and the MI approaches is presented in [73, Chapter 6].

2.3.3 Semi-Blind Source Separation (SBSS)

It is straightforward to apply BSS to MCAEC by treating the loudspeaker signals captured by microphones as some of the source signals to be separated as long as there are enough microphones [103]. However, the loudspeaker (*i.e.*, the far-end) signals are already known *a priori*, thus it is not necessary to perform the separation on them.

A better approach is to cast the MCAEC problem into one system framework consisting of BSS and AEC in terms of SBSS. Figure 9 illustrates a model for SBSS-based SAEC. The subscripts "n" and "f" denote "near end" and "far end," respectively. The far-end and the near-end linear mixing systems can be combined together systematically in the frequency domain as

$$\mathbf{H}(k,l) = \begin{bmatrix} \mathbf{H}_{11}(k,l) & \mathbf{H}_{12}(k,l)\mathbf{G}(k,l) \\ 0 & \mathbf{G}(k,l) \end{bmatrix} \tag{26}$$

for the k^{th} frequency bin and the l^{th} time block such that

$$\begin{bmatrix} \mathbf{x}_n(k,l) \\ \mathbf{x}_f(k,l) \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{11}(k,l) & \mathbf{H}_{12}(k,l)\mathbf{G}(k,l) \\ 0 & \mathbf{G}(k,l) \end{bmatrix} \begin{bmatrix} \mathbf{s}_n(k,l) \\ \mathbf{s}_f(k,l) \end{bmatrix}, \tag{27}$$

where

$$\begin{bmatrix} \mathbf{x}_n(k,l) \\ \mathbf{x}_f(k,l) \end{bmatrix} = \mathrm{DFT}\left\{ \begin{bmatrix} \mathbf{x}_n(n) \\ \mathbf{x}_f(n) \end{bmatrix} \right\} \tag{28}$$

is a 4-by-1 vector formed by the DFT of the near and far-end microphone signal vectors $\mathbf{x}_n(n) = [x_1(n), x_2(n)]^T$ and $\mathbf{x}_f(n) = [x_3(n), x_4(n)]^T$,

$$\begin{bmatrix} \mathbf{s}_n(k,l) \\ \mathbf{s}_f(k,l) \end{bmatrix} = \mathrm{DFT}\left\{ \begin{bmatrix} \mathbf{s}_n(n) \\ \mathbf{s}_f(n) \end{bmatrix} \right\} \tag{29}$$

is a 4-by-1 vector formed by the DFT of the near and far-end source signal vectors $\mathbf{s}_n(n) = [s_1(n), s_2(n)]^T$ and $\mathbf{s}_f(n) = [s_3(n), s_4(n)]^T$,

$$\mathbf{H}_{11}(k,l) = \mathrm{DFT}\{\mathbf{H}_{11}(n)\} = \mathrm{DFT}\left\{\begin{bmatrix} h_{11}(n) & h_{12}(n) \\ h_{21}(n) & h_{22}(n) \end{bmatrix}\right\} \tag{30}$$

is the 2-by-2 near-end response matrix between the near-end sources and microphones,

$$\mathbf{H}_{12}(k,l) = \mathrm{DFT}\{\mathbf{H}_{12}(n)\} = \mathrm{DFT}\left\{\begin{bmatrix} h_{13}(n) & h_{14}(n) \\ h_{23}(n) & h_{24}(n) \end{bmatrix}\right\} \tag{31}$$

is the 2-by-2 near-end response matrix between the near-end loudspeaker and microphones (*i.e.*, the echo paths), and

$$\mathbf{G}(k,l) = \mathrm{DFT}\{\mathbf{G}(n)\} = \mathrm{DFT}\left\{\begin{bmatrix} g_{11}(n) & g_{12}(n) \\ g_{21}(n) & g_{22}(n) \end{bmatrix}\right\} \tag{32}$$

is the 2-by-2 far-end response matrix.

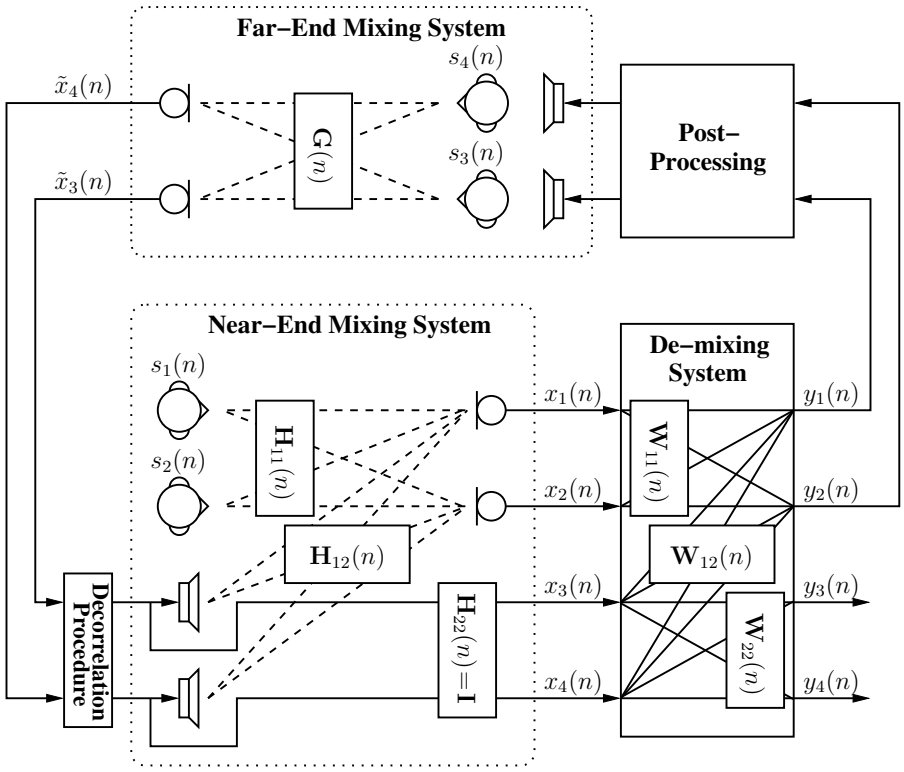


Figure 9: Model for SAEC based on semi-blind source separation (SBSS).

Then the objective is to find the 4-by-4 de-mixing matrix

$$\mathbf{W}(k,l) = \begin{bmatrix} \mathbf{W}_{11}(k,l) & \mathbf{W}_{12}(k,l) \\ 0 & \mathbf{W}_{22}(k,l) \end{bmatrix} \tag{33}$$

such that

$$\begin{bmatrix} \mathbf{y}_n(k,l) \\ \mathbf{y}_f(k,l) \end{bmatrix} = \begin{bmatrix} \mathbf{W}_{11}(k,l) & \mathbf{W}_{12}(k,l) \\ 0 & \mathbf{W}_{22}(k,l) \end{bmatrix} \begin{bmatrix} \mathbf{x}_n(k,l) \\ \mathbf{x}_f(k,l) \end{bmatrix} \simeq \begin{bmatrix} \mathbf{s}_n(k,l) \\ \mathbf{s}_f(k,l) \end{bmatrix}, \tag{34}$$

where $\mathbf{W}(k,l)$ can be estimated by using the NG algorithm of (25). The optimal solution is obtained when $\mathbf{W}(k,l)\mathbf{H}(k,l) = \mathbf{I}$, *i.e.*,

$$\mathbf{W}^{\text{opt}}(k,l)\mathbf{H}(k,l) = \begin{bmatrix} \mathbf{W}_{11}^{\text{opt}}(k,l)\mathbf{H}_{11}(k,l) & (\mathbf{W}_{11}^{\text{opt}}(k,l)\mathbf{H}_{12}(k,l) + \mathbf{W}_{12}^{\text{opt}}(k,l))\mathbf{G}(k,l) \\ 0 & \mathbf{W}_{22}^{\text{opt}}(k,l)\mathbf{G}(k,l) \end{bmatrix} = \mathbf{I}, \tag{35}$$

which requires that

$$(\mathbf{W}_{11}^{\text{opt}}(k,l)\mathbf{H}_{12}(k,l) + \mathbf{W}_{12}^{\text{opt}}(k,l))\mathbf{G}(k,l) = 0 \tag{36}$$

and

$$\mathbf{H}_{12}(k,l) = -\mathbf{W}_{11}^{\text{opt}}(k,l)^{-1}\mathbf{W}_{12}^{\text{opt}}(k,l) \tag{37}$$

if $\mathbf{G}(k,l)$ is fully ranked.

We can show through matrix manipulations that the sub-matrix $\mathbf{W}_{11}(k,l)$ is responsible for source separation while $\mathbf{W}_{12}(k,l)$ performs echo cancellation, where the two matrices are linked together with the near-end response matrix $\mathbf{H}_{12}(k,l)$, which represents the echo paths, through (37). More specifically, the BSS is performed such that $\mathbf{W}_{11}(k,l)\mathbf{H}_{11}(k,l)$ $\simeq \mathbf{I}$, and the estimation error from the AEC is obtained by $\mathbf{E}(k,l) = \mathbf{x}_n(k,l) - \mathbf{W}_{11}(k,l)^{-1}$ $\mathbf{W}_{12}(k,l)\mathbf{x}_f(k,l)$. (36) also indicates that just as in the traditional MCAEC framework, a change in the far-end room response $\mathbf{G}(k,l)$ disrupts the adaptation of the echo cancellation matrix $\mathbf{W}_{12}(k,l)$.

The SBSS framework was first proposed in [60] and applied successfully as a combination of BSS and single-channel AEC in [80] without double-talk detection, where such noise-robustness is naturally derived from the ICA optimization. We observed in [135] that

a decorrelation procedure on the far-end reference signals can indeed improve the SBSS performance for SAEC. On the other hand, we have also observed that a relatively high SAEC performance is still achievable without the decorrelation procedure if $\mathbf{W}_{22}(k,l)$ is constrained to be an identity matrix and the SBSS is implemented appropriately in a block-online fashion [90, 89].

There are many open issues yet to be solved for applying the ICA-based SBSS algorithm to real-time situations due to its computational complexity and slow convergence characteristics. Still, the approach is naturally suited for general signal enhancement purpose with multiple signals. The main difference of such an approach from the traditional MCAEC framework besides the use of the HOS over the SOS is that it is a truly multi-channel approach [89]. Although multiple loudspeakers and microphones are used, the conventional MCAEC is inherently a single-channel approach and thus is ill-suited when there are many interfering signals, some of which are often not directly observable, *e.g.*, local speech and background noise.

2.4 Effect of Distortions on AEC and BSS

2.4.1 Linear Distortion

An important question is whether or not tERLE > *a priori* SNR is possible after the convergence of the LMS algorithm, *i.e.*, $E\{|\bar{e}(n)|^2\} < E\{|v(n)|^2\}$ as $n \to \infty$. The steady-state convergence analysis in terms of the system distance for the regularized NLMS algorithm $\mathbf{w}(n+1) = \mathbf{w}(n) + \mu e(n)\frac{\mathbf{x}(n)}{\|\mathbf{x}(n)\|^2+\delta}$, where $\delta > 0$ is the regularization parameter, is summarized as follows [46]. Given $\sigma_x^2 = E\{|x(n)|^2\}$ and $\sigma_v^2 = E\{|v(n)|^2\}$,

- Convergence without step-size control or regularization:

$$\lim_{n\to\infty} E\{\|\mathbf{h}_\Delta(n)\|^2\}\big|_{\mu=1,\delta=0} \approx \frac{\sigma_v^2}{\sigma_x^2}. \tag{38}$$

- Convergence with step-size control only:

$$\lim_{n\to\infty} E\{\|\mathbf{h}_\Delta(n)\|^2\}\big|_{\delta=0} \approx \frac{\mu}{2-\mu}\frac{\sigma_v^2}{\sigma_x^2}. \tag{39}$$

- Convergence with regularization only:

$$\lim_{n\to\infty} E\{\|\mathbf{h}_\Delta(n)\|^2\}\big|_{\mu=1} \approx \frac{L\sigma_x^2}{L\sigma_x^2+2\delta}\frac{\sigma_v^2}{\sigma_x^2}. \tag{40}$$

Thus since $E\{|\bar{e}(n)|^2\} \approx \sigma_x^2 E\{\|\mathbf{h}_\Delta(n)\|^2\}$ according to (8) for white $x(n)$ and uncorrelated $x(n)$ and $h(n)$, achieving the echo cancellation below the average noise level is generally possible as long as there is either a step-size control or a regularization procedure applied to the NLMS algorithm.

We note that the above analysis holds when assuming white signals. In reality for a highly correlated signal such as speech in a complex acoustic environment, low misalignment is only a sufficient and not a necessary condition for high ERLE or tERLE. Misalignment determined in the time domain as defined by (15) is purely a measure of the system identification performance. A meaningful measure for the overall cancellation performance should practically be the tERLE for the linear distortion case and the ERLE for the nonlinear distortion case, and the two measures do not have to directly correlate with the MSD performance.

2.4.2 Nonlinear Distortion

Figure 10 illustrates the amount of distortion created by the GSM AMR speech codec [58] on a female speech. The figure shows that a speech coding distortion is directly proportional to the signal level and the bit-rate and that it starts to increase exponentially beyond 20 dB signal attenuation. The figure also indicates that the distortion is quite significant even if it is perceptually unnoticeable.

Figure 11 was obtained from simulating the loudspeaker saturation by using a saturation-type nonlinearity modeled by

$$\varphi(x) = \frac{1 - \exp(-x/\rho)}{1 + \exp(-x/\rho)}, \quad -1 \leq x \leq 1 \tag{41}$$

for input signal x and some saturation parameter $\rho > 0$, *i.e.*, smaller ρ gives more compression. The figure shows that while a mild loudspeaker nonlinearity would cause the ERLE to go down significantly, it is a speech coding distortion on the acoustic echo signal that ultimately limits the achievable ERLE.

Results from other simulations we performed in [134] showed that while the NLMS algorithm exhibits robustness against a saturation-type nonlinearity at low loudspeaker volume level, block-based algorithms such as FBLMS and RLS are unable by themselves

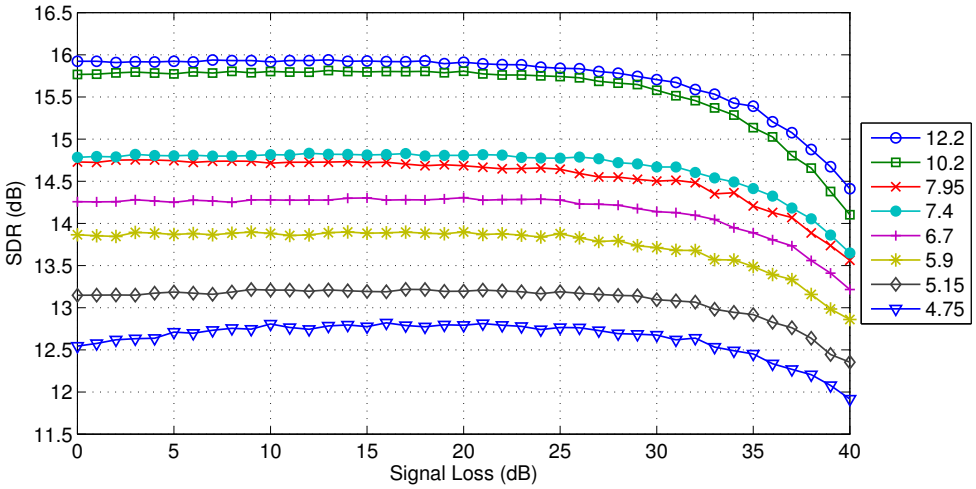


Figure 10: GSM AMR speech coding distortion, measured in terms of the signal-to-distortion ratio (SDR), versus input speech magnitude, measured in terms of the signal loss, for various bit-rate (kbps).

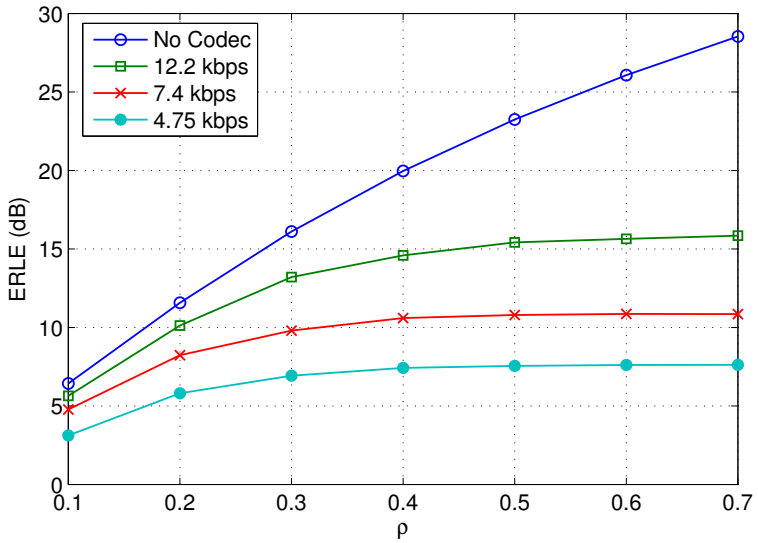


Figure 11: Echo return loss enhancement (ERLE) from the network-based AEC (using the FBLMS algorithm) as a function of the loudspeaker saturation parameter ρ (smaller ρ indicates greater saturation) and the GSM AMR bit-rate.

to adequately handle the effect of either the loudspeaker saturation or the speech coding distortion on the acoustic echo at any volume level. Figure 11, which was obtained from the FBLMS algorithm, not only indicates the severity of speech coding distortion but also the issues on how the nonlinearities should be treated with respect to an adaptive algorithm, *e.g.*, sample-based versus block-based, time-domain versus frequency-domain, polynomial versus memory-less modeling of nonlinearity.

Nonlinear AEC (NAEC) is used when the LEMS can be modeled by a Hammerstein system to decouple the echo path into a cascade of nonlinear response, *e.g.*, the loudspeaker saturation, and linear response, *i.e.*, the RIR itself. Main approaches to NAEC include using either a memory-less nonlinearity in the Wiener-Hammerstein cascaded model [122, 121] or a nonlinearity with memory, represented by the polynomial Volterra filter [123, 43] or by neural network [15], in a Hammerstein system to compensate for the nonlinear response. Other recent techniques consist of orthogonalized power filter with parallelized nonlinear modeling of the echo path [66], pre-distortion of the reference signal to linearize the echo path prior to the excitation of the LEMS [109], and optimal control of RES to remove the remaining nonlinear echo component after AEC [67, 108]. More detailed treatment of NAEC can be found in [105].

2.4.3 Sampling Rate Mismatch

We showed in [100] that a mismatch between two signals $x_1(n_1)$ and $x_2(n_2)$ with different sampling periods $T_1 \neq T_2$ may be corrected sufficiently by following a simple interpolation procedure in the time domain instead of the conventional approach of upsampling followed by downsampling. That is, if the signals are band-limited, $T_1 \approx T_2$, and the amount of mismatch is estimated accurately, the interpolation procedure using a truncated sinc function can be applied to $x_1(n_1)$ to match its sampling rate to that of $x_2(n_2)$, *i.e.*,

$$\hat{x}_1(n_2) \approx \sum_{n_1=-P}^{P} x_1(n_1)\,\mathrm{sinc}(t_1 - n_1), \tag{42}$$

where $2P+1$ is the filter order and $t_1 = n_2 T_2/T_1$ is the interpolated sample position in continuous time. The computational cost is further reduced by re-using the interpolation filter coefficients for a fixed $\check{t}_1 = t_1 - [t_1]$, where $[\cdot]$ is the greatest integer function (*i.e.*,

flooring function), to determine a block of interpolated samples for $\check{n}_2 \leq n_2 \leq \check{n}_2 + B - 1$, $\check{n}_2 = \check{t}_1 T_1 / T_2$.

Table 2 illustrates the variations in the actual sampling rate of portable audio capturing devices (cell phones, PDAs, laptops). The table shows that a mismatch can be over 0.5% of the nominal sampling rate. Figure 12 illustrates the effect of the sampling rate mismatch on AEC before and after the mismatch correction of (42) using $P = 64$ and various re-use block length B. White Gaussian noise (WGN) was added at 40 dB signal-to-noise ratio (SNR) to a simulated acoustic echo of 200 ms in length to create a more realistic acoustic mixing condition. The figure shows that a mere 100 ppm mismatch (0.01%) in the sampling rate can decrease the ERLE by 10 dB. Only about 5 dB ERLE would then be possible without the sampling rate correction for Device 5 and 6 in Table 2, for which the re-use block size has to be less than 64 to achieve around 20 dB ERLE.

Table 2: Sampling rate mismatch, measured in terms of parts per million (ppm), for six portable audio capturing devices with the nominal sampling rate of 16 kHz.

	Device 1	Device 2	Device 3	Device 4	Device 5	Device 6
Mismatch (ppm)	137	141	145	942	5860	-6140

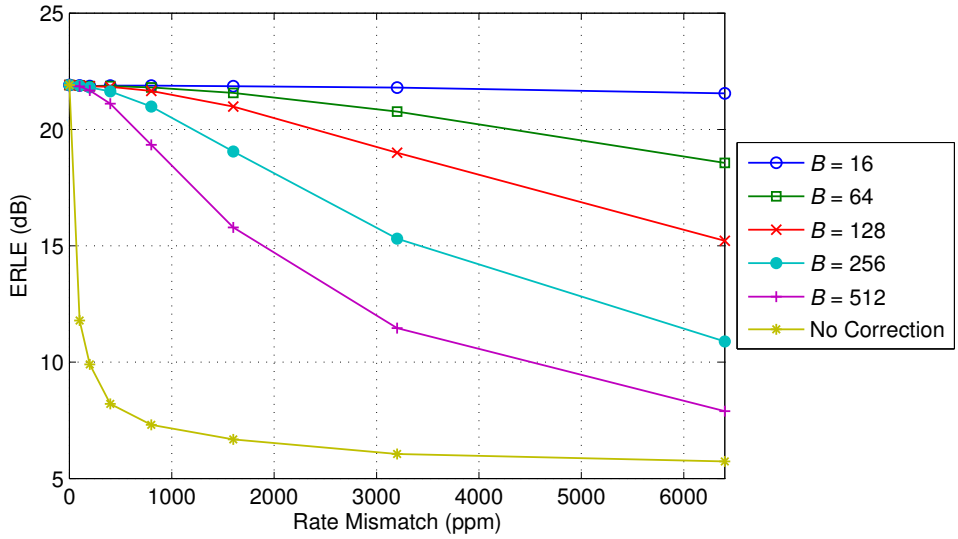


Figure 12: ERLE when there is a sampling rate mismatch between the reference signal and the acoustic echo. The correction was made through re-sampling by interpolation in the time domain by using the coefficients re-use block size of B.

CHAPTER III

ROBUST AEC VIA RESIDUAL ECHO ENHANCEMENT (REE)

The least mean square (LMS) algorithm is a popular choice for acoustic echo cancellation (AEC) due to its computational simplicity and adaptability. The filter error $e(n)$ is modeled to be quadratic in the space of the filter coefficients vector $\mathbf{w}(n)$, and the adaptation rule is based on the second-order statistics (SOS). The LMS algorithm has a natural ability to eventually converge to a unique and optimal solution that minimizes the mean square error (MSE) $E\{|e(n)|^2\}$ and consequently satisfies the orthogonality principle $E\{e(n)x(n)\} = 0$, where $x(n)$ is the reference signal. However, along with the gradient noise, local distortions (*e.g.*, double talk, speech codec) can greatly disrupt the LMS adaptation process and contribute to slow convergence and large MSE.

There is ultimately a need for a residual echo suppressor (RES) based on classical signal enhancement, namely the Wiener filtering, to attenuate the remaining cancellation error to an acceptable level [46], *i.e.*, the reduction of $\bar{e}$ within e. Similar in spirit to the joint statistical adaptation of AEC and RES studied in [45], we will motivate in this chapter the idea of *system approach* to achieving the robust signal enhancement performance. The overall goal is a proper integration of individually designed components, or algorithms, to mutually support one another and achieve the robustness of the entire AEC system rather than focusing simply on the AEC adaptive filtering component itself. Specifically, we postulate that if an adaptive filter is capable of converging to the optimal solution, then the adaptation process can be assisted if the effects of disruption are removed before the filter coefficients are updated, *i.e.*, the reduction of v within e prior to the filter adaptation. In other words, by including a signal enhancement procedure not just for the post-enhancement purpose to further reduce the residual echo but within the feedback-loop of an adaptive filter to *enhance* the error, the linear portion of the impulse response should be better estimated iteratively and recursively through a stochastic adaptation procedure.

We investigated in [133] the new "system" perspective to applying an instantaneous memoryless nonlinearity that performs the signal enhancement of the residual echo in order to increase the robustness of the AEC in a continuously noisy environment [128, 129, 130]. We will hereafter refer to such a nonlinearity as the error recovery nonlinearity (ERN). The procedure, represented schematically in Figure 13, is summarized by the LMS update equation

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu f(e(n))\mathbf{x}(n), \tag{43}$$

where the ERN $f(\cdot)$ is applied to the filter estimation error $e(n)$. We refer to such a procedure hereafter as residual echo enhancement (REE), or simply error enhancement in general. The use of an error nonlinearity to modify the convergence behavior of the LMS algorithm has been addressed on a number of occasions in the past: [117, 30, 76, 14, 104, 28, 2], just to name a few. Although the same signal-enhancement approach to "enhance" the residual echo was recently studied also by others in [114], we are to the best of our knowledge the first to propose and demonstrate the effectiveness of the technique for AEC in [128, 129, 130]. The idea is akin to the Bussang technique for blind equalization [48], where the system inversion without the reference signal is made possible through the iterative application of a linear adaptive filter, which deconvolutes the observed signal, followed by a memoryless nonlinearity, which reduces the modeling error (referred to as the "convolutional" noise in [48]) to assist the inversion-adaptation process.

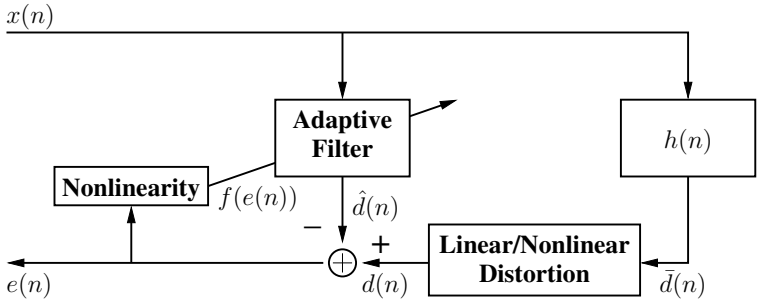


Figure 13: Adaptive filtering with linear or nonlinear distortion on the true, noise-free response $\bar{d}(n)$. A noise-suppressing memoryless nonlinearity $f(\cdot)$ is applied to the observed filter estimation error $e(n)$ in the feedback-loop in order to suppress the effects of the distortion during the adaptation of a linear filter.

We have observed that introducing the noise-suppressing nonlinearity to the residual

error in such a manner is similar to minimizing the MSE or the mean square deviation (MSD) $E\{\|\mathbf{h}_\Delta(n)\|^2\}$, where $\mathbf{h}_\Delta(n)$ is the misalignment vector, of LMS-type algorithms. Furthermore, the ERN can also be viewed as a function that controls the step-size μ for sub-optimal conditions reflected in the error statistics when the signals are no longer Gaussian distributed, as most often is the case in reality, and it may be combined with other existing noise-robust schemes to improve the overall performance of the LMS algorithm. More importantly, we have shown in [131] that the technique is well-founded in an information-theoretic sense and has a deep connection to the algorithms based on independent component analysis (ICA) [56], *e.g.*, blind source separation (BSS) that maximizes the independence of output signals and allows the recovery of a target signal among interfering signals even when the source signals are unknown. In fact, the single-channel AEC problem is a special case of semi-blind source separation (SBSS) for which one of the source signals is partially known, *i.e.*, the far-end reference signal that becomes the acoustic echo. With such a built-in noise-robust mechanism, continuous adaptation is permitted during double talk by a proper combination of the ERN and the LMS algorithm.

The error enhancement paradigm is underlined by the effectiveness of taking a broader system view of the signal enhancement problem in general. Often times the constrained focus on a specific algorithm, governed by mathematics, leads to limited applicability in the physical world. By placing all possible signals from both the near and the far end, their prior source information, and the system components together into a proper perspective, robust performance is achieved in a very complex, real-world setting with non-ideal acoustic mixing conditions. Moreover, if adapted appropriately, the system approach alleviates the need for precise tracking of the room impulse response (RIR) or the signal statistics to obtain sufficient performance. In particular, we introduce what we refer to as the *block-iterative adaptation* (BIA), which is related to the traditional data reuse and batch adaptation procedures. The system perspective is already exemplified by the ICA-based frequency-domain BSS systems developed previously by many others and will be illustrated here in the scope of single-channel AEC. Extension of the principle of orthogonality to the maximization of the independence between the involved signals leads naturally to the ERN and offers a

refreshingly new view of the conventional AEC problem. Our proposed approach in this chapter brings seemingly separate signal enhancement techniques under one roof and ties them together in an analytically consistent and practical manner.

The rest of this chapter is organized as follows. First in Section 3.1, we review the conventional approaches to robustness during LMS-based AEC. Next in Section 3.2, we derive and analyze the ERN for REE in detail; show the REE technique's connections to the traditional AEC procedures, ICA, and SBSS; and discuss the system approach to the AEC problem and the BIA procedure. Finally in Section 3.3, we provide the experimental results from REE when dealing with linear and nonlinear distortions.

3.1 Conventional Approaches

3.1.1 Robustness against Non-stationarity

The normalized least mean square (NLMS) algorithm [139] is preferred over the LMS algorithm since the real-world acoustic signals involved in communications are rarely stationary, which leads to the gradient noise during the LMS adaptation. The algorithm can be derived through the *principle of minimal disturbance*, which states that "from one iteration to the next, the weight vector of an adaptive filter should be changed in a minimal manner, subject to a constraint imposed on the updated filter's output" [49].

In the NLMS algorithm, the LMS filter coefficients adaptation rule is "normalized" by dividing the reference signal by an estimate of its power:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu e(n) \frac{\mathbf{x}(n)}{\|\mathbf{x}(n)\|^2}, \tag{44}$$

where $\|\cdot\|$ is the Euclidian norm. By such a normalization procedure, the NLMS algorithm becomes invariant to the scaling of the reference signal, and its convergence $\|\mathbf{w}(n+1) - \mathbf{w}(n)\| \to \epsilon$ as $n \to \infty$ for arbitrarily small $\epsilon > 0$ is guaranteed for $0 < \mu < 2$.

3.1.2 Robustness against Ambient Noise

Just as with the LMS algorithm, the NLMS algorithm is still susceptible to the effect of distortions on the acoustic echo. A simple procedure for improving the convergence behavior in the presence of a local additive noise v is to normalize the reference signal vector $\mathbf{x}$ with

$\|\mathbf{x}\|^2 + \delta$ for $\delta > 0$ [46] to regularize, or stabilize, the adaptation when $\|\mathbf{x}\| \approx 0$:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu e(n) \frac{\mathbf{x}(n)}{\|\mathbf{x}(n)\|^2 + \delta}. \tag{45}$$

While the regularization parameter δ can be maintained large enough to keep the NLMS algorithm from diverging when the signal-to-noise ratio (SNR) between x and v is very small, there is no explicit guidance on how to control the parameter precisely in practical situations with the time-varying SNR.

Many adaptive, or variable, step-size algorithms were developed not only to find the optimal step-size μ to improve the stability of the LMS and the NLMS algorithms in the presence of v, *e.g.*, [13, 77, 26, 1, 71, 5, 65, 74], but also to specifically reduce the effect of v on the adaptive algorithms, *e.g.*, [52, 97, 91, 50, 124]. The first set of approaches keeps μ as large as possible during early stages of adaptation for fast convergence and reduces it when the true error $\bar{e}$ or the misalignment becomes small as the algorithm converges and the effect of v increases. The second set of approaches generally controls μ as a function of the signal statistics, where μ is adjusted to be large at high SNR and small at low SNR.

3.1.3 Robustness against Double-Talk

Most of the adaptive step-size techniques listed above except for [124] are designed for ambient and continuous background noise. They do not work as well during the double talk situation when the local noise is dominated by a near-end speech signal, which can be highly non-stationary and "bursty" (*i.e.*, occurs suddenly with large energy). There are frequency-domain adaptive step-size algorithms designed for dealing with double talk [112, 126, 127, 114]. Still, distinguishing the near-end speech from the far-end speech in the observed, noisy acoustic echo in order to estimate the SNR is a very complicated task, especially when only a single-channel recording is available.

A traditional and practical solution is to stop the adaptation completely during double talk by invoking a double-talk detector (DTD), in which a decision is made by making a comparison between the available signals, *e.g.*, the reference signal (the far-end microphone signal), the near-end microphone signal, the estimated acoustic echo, and the residual echo

in terms of the magnitude ratio [29] (the Geigel algorithm), the cross correlation (or coherence) [145, 39, 10, 35, 61, 37, 59], or the mutual information [107, 106]. The two-path models approach, which uses a "background" filter only for adaptation and a "foreground" filter for actual cancellation, can also be utilized to overcome the sporadic disturbances from a local speech without entirely freezing the adaptation process [92, 111, 25]..

A more advanced approach is to apply a compressive, *i.e.*, saturation-type, memoryless nonlinearity to the estimation error to limit the sudden jump in the error that may occur when a DTD fails to detect the event initially [33, 38, 7, 18, 113]. The technique is based on the robust statistics theory [54], which provides the means to minimize the influence of the outliers on a statistical estimation procedure. There is a trade-off between the false alarm, which unduly stops the adaptation and retards the convergence rate, and the missed detection, which may lead to the adaptation instability. A logical solution for handling the two-sided problem is to adjust the DTD threshold level accordingly to decrease the false alarm rate at the cost of increased missed detections, in which case the effect of the double-talk leakage is then limited by a compressive error nonlinearity.

3.2 Residual Echo Enhancement

3.2.1 Error Recovery Nonlinearity (ERN)

3.2.1.1 Bussang Technique for Unsupervised Adaptation

Residual echo enhancement closely follows the Bussang technique for blind equalization or deconvolution:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu(z(n) - \phi(z(n)))\mathbf{y}(n), \tag{46}$$

where $\mathbf{y}(n)$ is a vector of the observed (convoluted) signal $y(n) = h(n) * x(n)$, $z(n) = w(n) * y(n) \approx \mathbf{w}^T(n)\mathbf{y}(n)$ is the filtered (deconvoluted) output, and $\phi(\cdot)$ is a memoryless nonlinearity such that $\phi(z(n)) = \hat{x}(n)$ is an estimator of the original, unobservable signal $x(n)$. The system equalization without any training signal is made possible through the iterative application of a linear adaptive filter, which estimates the inverse filter $\mathbf{w}(n)$, followed by the nonlinearity, which enhances the filtered signal $z(n)$. Specifically, a decomposition

of $z(n)$ reveals that

$$\begin{aligned} z(n) &= w(n) * h(n) * x(n) \\ &= x(n) - h^{-1}(n) * h(n) * x(n) + w(n) * h(n) * x(n) \\ &= x(n) + v(n), \end{aligned} \tag{47}$$

where $h^{-1}(n)$ is the inverse impulse response, *i.e.*, $h^{-1}(n) * h(n) = 1$, and $v(n) = (w(n) - h^{-1}(n)) * h(n) * x(n)$ is the convolutional noise. Hence by reducing the effect of $v(n)$ in $z(n)$ through the application of the nonlinearity $\phi(\cdot)$, the estimation error $e(n) = z(n) - \phi(z(n)) = z(n) - \hat{x}(n)$ necessary for the filter adaptation can be obtained blindly.

3.2.1.2 Derivation of Noise-Suppressing Nonlinearities

Assuming the noise $v(n)$ is statistically independent from the true estimation error $\bar{e}(n)$, the objective is to derive an optimal instantaneous nonlinearity that recovers as much of $\bar{e}(n)$ as possible from the observed noisy error $e(n) = \bar{e}(n) + v(n)$. Hereafter, the sample-wise signal enhancement is assumed, and the time index n may be dropped for simplification in notation.

As done commonly in the signal enhancement community and also in [48] for the Bussang technique, the Bayesian parameter estimation procedure [99] is utilized to perform the error enhancement. Namely, the best estimate of $\bar{e}$ can be found through either the minimum mean-square error (MMSE) or the maximum a-posteriori probability (MAP) estimation procedure [1] (the MAP estimate may also be referred to as the maximum likelihood (ML) estimate [42, 55] [2]). At the heart of the Bayesian estimation is the conditional probability

[1] For the random variables that consist of the unknown parameter θ, the observation y, and the estimate $\hat{\theta}(y)$, the MMSE procedure minimizes the risk $r(\hat{\theta}) = E_\theta\{C[\hat{\theta}(y), \theta]\}$ for the quadratic cost $C[\hat{\theta}, \theta] = (\hat{\theta} - \theta)^2$, whereas the MAP procedure minimizes the risk for the uniform cost $C[\hat{\theta}, \theta] = 1$, if $|\hat{\theta} - \theta| > \Delta$ for $\Delta > 0$, and $C[\hat{\theta}, \theta] = 0$, otherwise. There is another approach not covered here called the minimum mean-absolute error (MMAE) procedure that minimizes the risk for the absolute cost $C[\hat{\theta}, \theta] = |\hat{\theta} - \theta|$.

[2] Generally in statistics, the ML estimate is defined to be a special case of the MAP estimate with the prior probability $p(\theta) = 1$ such that it maximizes the likelihood $p(y|\theta)$. The ML procedure defined as such is not covered here since it gives the trivial solution $\hat{\theta} = y$ when used as a sample-wise estimator. Some procedures may be referred to as being ML instead of MAP even though the prior probability is not explicitly set equal to 1 since they still maximize the joint density (or "likelihood") $p(y, \theta) = p(y|\theta)p(\theta)$.

given by the Bayes formula

$$p_{\bar{e}|e}(\bar{e}|e) = \frac{p_{e|\bar{e}}(e|\bar{e})p_{\bar{e}}(\bar{e})}{\int_{-\infty}^{\infty} p_{e|\bar{e}}(e|\bar{e})p_{\bar{e}}(\bar{e})d\bar{e}}. \tag{48}$$

The MMSE estimate is obtained by minimizing the expectation of the residual $E\{(\bar{e}-\hat{\bar{e}})^2|e\}$ with respect to the estimate $\hat{\bar{e}}$ conditioned on the observation e, resulting in $\hat{\bar{e}} = E\{\bar{e}|e\}$:

$$f_{\text{MMSE}}(e) = \int_{-\infty}^{\infty} \bar{e}\ p_{\bar{e}|e}(\bar{e}|e)d\bar{e}. \tag{49}$$

The MAP estimate is obtained by maximizing (48) directly, which is equivalent to maximizing the numerator only:

$$f_{\text{MAP}}(e) = \underset{\bar{e}}{\text{argmax}}\, p_{e|\bar{e}}(e|\bar{e})p_{\bar{e}}(\bar{e}). \tag{50}$$

To begin with, let the score function $\phi_s(\cdot)$ be defined as

$$\phi_s(s) = -\frac{\partial}{\partial s}\log p_s(s) = -\frac{p_s'(s)}{p_s(s)} \tag{51}$$

for some probability density function (PDF) p_s, where "$'$" is the derivative operator. (51) measures the relative rate at which p_s changes at a value s. Let us consider three random variables s, t, and u, where $s = t + u$ to reflect the additive distortion model. If the PDF of t is zero-mean Gaussian

$$p_t(t;\sigma_t) = \frac{1}{\sqrt{2\pi\sigma_t^2}}\exp\left(\frac{-t^2}{2\sigma_t^2}\right), \tag{52}$$

denoted hereafter by $t \sim Gaussian(0, \sigma_t)$, then

$$\phi_t(t;\sigma_t) = t/\sigma_t^2. \tag{53}$$

If u is from any probability distribution, then it can be shown by the convolution theorem for the PDF of the sum of independent random variables that (see Appendix A)

$$\phi_{s(u)}(s;\sigma_t) = \frac{1}{\sigma_t^2}\frac{\int_{-\infty}^{\infty} t\ p_u(s-t)\exp\left(\frac{-t^2}{2\sigma_t^2}\right)dt}{\int_{-\infty}^{\infty} p_u(s-t)\exp\left(\frac{-t^2}{2\sigma_t^2}\right)dt}, \tag{54}$$

where "$s;\sigma_t$" and "$s(u)$" indicate explicit dependence of s on t (which is integrated out) and u (which remains random), respectively. Specifically, if the PDF of u is zero-mean Laplacian

$$p_u(u;\alpha_u) = \frac{1}{2\alpha_u}\exp\left(\frac{-|u|}{\alpha_u}\right), \tag{55}$$

denoted hereafter by $u \sim Laplacian(0, \alpha_u)$, then

$$\phi_u(u; \alpha_u) = \text{sign}(u)/\alpha_u, \tag{56}$$

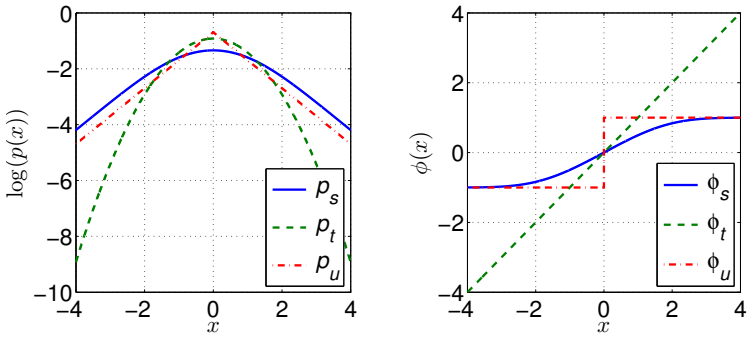
and the integration-by-parts rule gives

$$p_s(s; \sigma_t, \alpha_u) = \frac{\exp\left(\frac{-\chi^2}{2}\right)}{4\alpha_u}\left[\text{erfcx}\left(\frac{\xi - \psi}{\sqrt{2\xi}}\right) + \text{erfcx}\left(\frac{\xi + \psi}{\sqrt{2\xi}}\right)\right], \tag{57}$$

where $\chi = s/\sigma_t$ and $\psi = s/\alpha_u$ are the scaled signals, $\xi = \sigma_t^2/\alpha_u^2$ is the SNR (in a general sense), $\text{erfcx}(r) = \exp(r^2)\text{erfc}(r)$ is the scaled complementary error function, and $\text{erfc}(r) = \frac{2}{\sqrt{\pi}}\int_r^\infty \exp(-\bar{r}^2)d\bar{r}$ is the complementary error function. (57) then leads to

$$\phi_s(s; \sigma_t, \alpha_u) = \frac{1}{\alpha_u}\left[\frac{\text{erfcx}\left(\frac{\xi-\psi}{\sqrt{2\xi}}\right) - \text{erfcx}\left(\frac{\xi+\psi}{\sqrt{2\xi}}\right)}{\text{erfcx}\left(\frac{\xi-\psi}{\sqrt{2\xi}}\right) + \text{erfcx}\left(\frac{\xi+\psi}{\sqrt{2\xi}}\right)}\right]. \tag{58}$$

The PDFs (52), (55), (57), and the score functions (53), (56), (58), are plotted in Figure 14 for $\sigma_t = \alpha_u = 1$.



(a) PDF. (b) Score function.

Figure 14: PDFs (scaled by $\log(\cdot)$) and score functions of s, t, and u when $s = t + u$, $t \sim Gaussian(0, 1)$, and $u \sim Laplacian(0, 1)$. For $|x| < 2$, $\phi_s(x)$ exhibits approximately linear scaling by a factor of 0.5 that accounts for relatively equal contributions from t and u for small observed (noisy) magnitude $|s|$. For $|x| > 2$, $\phi_s(x) \to \phi_u(x)$ very rapidly as $|x| \to \infty$ since the kurtosis is higher for u than t, *i.e.*, the probability of u is greater than that of t for large observed magnitude $|s|$.

Specific forms of the ERN obtained from the MMSE and the MAP estimations are as follows (see Appendix A for the derivation). First, if $\bar{e} \sim Gaussian(0, \sigma_{\bar{e}})$ and v is any random variable, then

$$f_{\text{MMSE}}^{\text{GA}}(e) = \sigma_{\bar{e}}^2 \phi_{e(v)}(e; \sigma_{\bar{e}}), \tag{59}$$

$$f_{\text{MAP}}^{\text{GA}}(e) = \{\hat{\bar{e}} : \hat{\bar{e}} = \sigma_{\bar{e}}^2 \phi_v(e - \hat{\bar{e}})\}, \tag{60}$$

where the superscripts "G" and "A" indicate "Gaussian" error $\bar{e}$ and "any" type of distortion v, respectively (such a notation will be used hereafter). Specifically for $v \sim \mathit{Laplacian}(0, \alpha_v)$,

$$f_{\text{MMSE}}^{\text{GL}}(e) = \sigma_{\bar{e}}^2 \phi_e(e; \sigma_{\bar{e}}, \alpha_v). \tag{61}$$

$$f_{\text{MAP}}^{\text{GL}}(e) = \begin{cases} \text{sign}(e)\sigma_{\bar{e}}^2/\alpha_v & \text{if } |e| \geq \sigma_{\bar{e}}^2/\alpha_v, \\ e & \text{if } |e| < \sigma_{\bar{e}}^2/\alpha_v. \end{cases} \tag{62}$$

Next, if $v \sim \mathit{Gaussian}(0, \sigma_v)$ and $\bar{e}$ is any random variable, then

$$f_{\text{MMSE}}^{\text{AG}}(e) = e - \sigma_v^2 \phi_{e(\bar{e})}(e; \sigma_v), \tag{63}$$

$$f_{\text{MAP}}^{\text{AG}}(e) = \{\hat{\bar{e}} : \hat{\bar{e}} = e - \sigma_v^2 \phi_{\bar{e}}(\hat{\bar{e}})\}. \tag{64}$$

Specifically for $\bar{e} \sim \mathit{Laplacian}(0, \alpha_{\bar{e}})$,

$$f_{\text{MMSE}}^{\text{LG}}(e) = e - \sigma_v^2 \phi_e(e; \sigma_v, \alpha_{\bar{e}}), \tag{65}$$

$$f_{\text{MAP}}^{\text{LG}}(e) = \begin{cases} e - \text{sign}(e)\sigma_v^2/\alpha_{\bar{e}} & \text{if } |e| \geq \sigma_v^2/\alpha_{\bar{e}}, \\ 0 & \text{if } |e| < \sigma_v^2/\alpha_{\bar{e}}. \end{cases} \tag{66}$$

Finally, if $\bar{e} \sim \mathit{Gaussian}(0, \sigma_{\bar{e}})$ and $v \sim \mathit{Gaussian}(0, \sigma_v)$, then

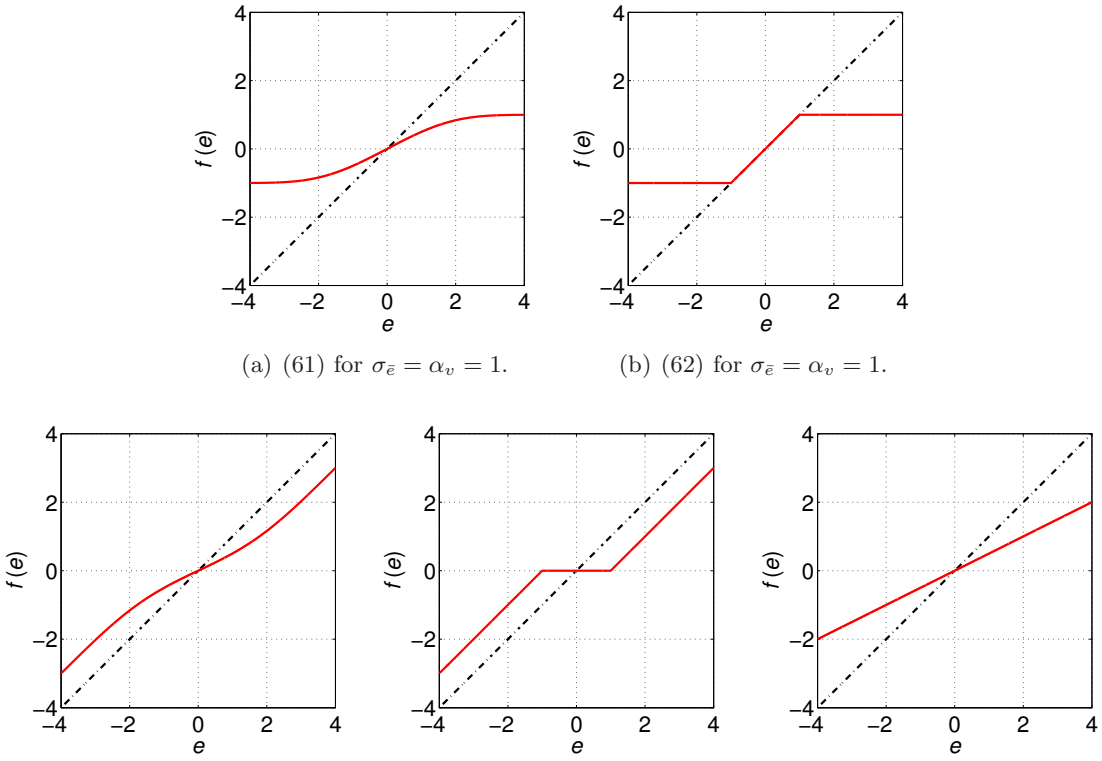
$$f_{\text{MMSE}}^{\text{GG}}(e) = f_{\text{MAP}}^{\text{GG}}(e) = \frac{\sigma_{\bar{e}}^2}{\sigma_{\bar{e}}^2 + \sigma_v^2} e, \tag{67}$$

which is the well-known Wiener scaling rule used often by the traditional frequency-domain signal enhancement techniques. The nonlinearities (61), (62), (65), (66), and (67) are plotted for $\sigma_{\bar{e}} = \alpha_v = 1$, $\alpha_{\bar{e}} = \sigma_v = 1$, and $\sigma_{\bar{e}} = \sigma_v = 1$, respectively, in Figure 15.

3.2.1.3 Ad-Hoc Nonlinearities

The center-clipping (CC) and coring (COR) nonlinearities, used widely for image and audio signal enhancements, can be defined as

$$f_{\text{CC}}(e) = \begin{cases} e & \text{if } |e| \geq k_{cc}, \\ 0 & \text{if } |e| < k_{cc}, \end{cases} \tag{68}$$



(a) (61) for $\sigma_{\bar{e}} = \alpha_v = 1$.

(b) (62) for $\sigma_{\bar{e}} = \alpha_v = 1$.

(c) (65) for $\alpha_{\bar{e}} = \sigma_v = 1$.

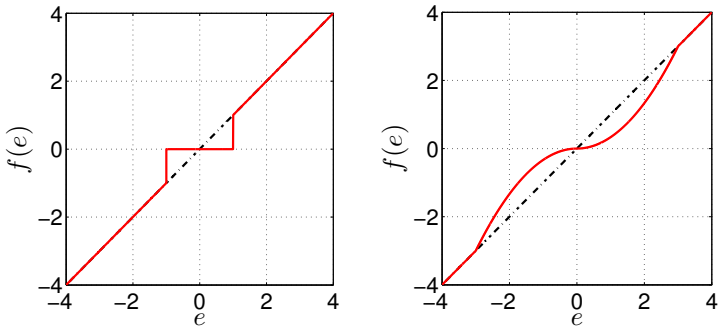
(d) (66) for $\alpha_{\bar{e}} = \sigma_v = 1$.

(e) (67) for $\sigma_{\bar{e}} = \sigma_v = 1$.

Figure 15: Realization of noise-suppressing nonlinearities (61), (62), (65), (66), and (67) obtained through the MMSE and the MAP estimation procedures when the observed adaptive filter estimation error is modeled additively as $e = \bar{e} + v$.

$$f_{\mathrm{COR}}(e) = \begin{cases} e & \text{if } |e| \geq k_{cor}, \\ \mathrm{sign}(e)\,|e/k_{cor}|^{\eta}\,k_{cor} & \text{if } |e| < k_{cor}, \end{cases} \tag{69}$$

for $k_{cc} > 0$, $k_{cor} > 0$, and $\eta \geq 1$. (68) and (69) are plotted in Figure 16(a) and 16(b), respectively, for $k_{cc} = 1$, $k_{cor} = 3$, and $\eta = 2$.



(a) (68) for $k_{cc} = 1$. (b) (69) for $k_{cor} = 3$, $\eta = 2$.

Figure 16: Realization of ad-hoc noise-suppressing nonlinearities (68) and (69).

3.2.1.4 Relationship between MMSE and MAP Nonlinearities

(67) can be interpreted as the "mid-point" of (59) and (63), where (59) $\rightarrow$ (61) and (63) $\rightarrow$ (65) when p_v for (59) and $p_{\bar{e}}$ for (63) become Laplacian-like, *i.e.*, has higher kurtosis than the Gaussian PDF. For example, Figure 14 illustrates that when the kurtosis is higher for v than $\bar{e}$, ϕ_e tends to represent ϕ_v for large $|e|$ at which the probability of v is higher than that of $\bar{e}$, and re-scaling the output of ϕ_e by $\sigma_{\bar{e}}^2$ results in the best estimate of $\bar{e}$. On the other hand, the opposite effect takes place for (65) when the kurtosis is higher for $\bar{e}$ than v, in which case ϕ_e is more likely to represent $\phi_{\bar{e}}$ instead of ϕ_v, and re-scaling the output of ϕ_e by σ_v^2 and then subtracting the result from e consequently gives the best estimate of $\bar{e}$. Using the Laplacian PDF for both $\bar{e}$ and v gives the nonlinearity similar in form to (61) if $\alpha_{\bar{e}} < \alpha_v$ and to (65) if $\alpha_{\bar{e}} > \alpha_v$, whereas the nonlinearity becomes identical to the linear Wiener scaling of (67) if $\alpha_{\bar{e}} = \alpha_v$ [75]. Whether or not the overall form is "compressive" as in (61) rather than "subtractive" (via spectral subtraction) as in (65) is determined by the score function ϕ_e that depends on the degree of non-Gaussianity of $p_{\bar{e}}$ or p_v and on the SNR.

Similar analysis applies to the MAP nonlinearities (60), (62), (64), and (66), which are the limits of the MMSE nonlinearities, *i.e.*, (61) $\rightarrow$ (62) and (65) $\rightarrow$ (66) asymptotically as $|e| \rightarrow 0$ or ∞. Thus (61) and (65) may be approximated by (62) and (66), respectively, for large $|e|$ at which the MMSE nonlinearities are numerically unstable due to the division of p_e' by p_e. Moreover, (59) $\rightarrow$ (60) and (63) $\rightarrow$ (64) as $n \rightarrow \infty$ since $\phi_e(e) \rightarrow \phi_v(e)$ as the LMS algorithm dynamically reduces $\bar{e}$ and since the estimates $\hat{\bar{e}}$ and $\hat{v}$ are related by $e = \hat{\bar{e}} + \hat{v}$. The limiting behavior is consistent with the previous observation that the general form of the nonlinearity is governed by the Gaussianity of $p_{\bar{e}}$ and p_v. Also, it implies that the overall amount of scaling performed by the nonlinearity depends ultimately on the SNR. For example, the effect of (62) can be interpreted as multiplying $\sigma_{\bar{e}}$ by $\sigma_{\bar{e}}/\alpha_v$ to get the best estimate of $\bar{e}$ above the threshold $\sigma_{\bar{e}}^2/\alpha_v$, whereas the same interpretation holds for (66), in which case the ratio of and the threshold of interest are $\sigma_v/\alpha_{\bar{e}}$ and $\sigma_v^2/\alpha_{\bar{e}}$, respectively.

Since (67) is a linear estimator of $\bar{e}(n)$ by definition is not a nonlinearity. However, for a constant σ_v^2, the effect of (67) is to scale down $e(n)$ when $\bar{e}(n)$ is small (*i.e.*, $\sigma_{\bar{e}}^2/(\sigma_{\bar{e}}^2+\sigma_v^2) \rightarrow 0$ as $\sigma_{\bar{e}}^2 \rightarrow 0$) and leave $e(n)$ as it is when $\bar{e}(n)$ is large (*i.e.*, $\sigma_{\bar{e}}^2/(\sigma_{\bar{e}}^2 + \sigma_v^2) \rightarrow 1$ as $\sigma_{\bar{e}}^2 \rightarrow \infty$). Thus (67) dynamically emulates a coring nonlinearity for the case of ambient background noise [128].

3.2.2 Connection to Conventional Approaches

3.2.2.1 Optimal Error Nonlinearity

When the observed adaptive filter error is modeled as $e = \bar{e} + v$, where $\bar{e}$ is the true, zero-mean Gaussian-distributed error, the optimal error nonlinearity for the LMS algorithm that minimizes the steady-state MSE is [2]

$$f_{\text{LMS}}^{\text{opt}}(e) = -\frac{p_e'(e)}{p_e(e)} = \phi_e(e), \tag{70}$$

which is simply the score function (51) in terms of e. Since (70) is equal to (59) for zero-mean standardized Gaussian-distributed $\bar{e}$ (*i.e.*, $\sigma_{\bar{e}} = 1$) or $\sigma_{\bar{e}}^2$ in (59) may be absorbed into the step-size μ of the LMS algorithm, such an ERN is optimal in the LMS-sense for any distribution of the local noise v.

3.2.2.2 Adaptive Step-Size Procedure

The optimal adaptive step-size for the NLMS algorithm that achieves the largest decrease in the MSD is [71]

$$\mu_{\text{opt}}(n) = \frac{E\{|\bar{e}(n)|^2\}}{E\{|e(n)|^2\}} = \frac{E\{|\bar{e}(n)|^2\}}{E\{|\bar{e}(n)|^2\} + E\{|v(n)|^2\}} \tag{71}$$

when $\bar{e}$ and v are uncorrelated. (71) is precisely the Wiener scaling coefficient in (67), *i.e.*, the MMSE and the MAP estimation procedures provide the MSD-optimal step-size for the NLMS algorithm with $\mu = 1$ when $\bar{e} \sim Gaussian(0, \sigma_{\bar{e}})$ and $v \sim Gaussian(0, \sigma_v)$. The application of (67) to the NLMS algorithm means that while the reference signal vector $\mathbf{x}(n)$ is normalized to give a proper direction of adaptation in the solution space of $\mathbf{h}(n)$, the observed estimation error $e(n)$ is re-scaled to be nearly equal to $E\{|\bar{e}|\}$ to provide as correct amount of update as possible for $\mathbf{w}(n)$.

3.2.2.3 Regularization Procedure

A very effective adaptive regularization procedure for the NLMS algorithm proposed in [52] and utilized in [128, 129, 130, 131] is according to

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu e(n) \frac{\|\mathbf{x}(n)\|^2}{\|\mathbf{x}(n)\|^4 + \gamma \sigma_v^4} \mathbf{x}(n) \tag{72}$$

for $\gamma > 0$. The regularized normalization factor in (72) can be expanded into the product of a non-regularized normalization factor and a Wiener-like scaling factor [131]:

$$\frac{||\mathbf{x}(n)||^2}{||\mathbf{x}(n)||^4 + \gamma \sigma_v^4} \approx \frac{1}{||\mathbf{x}(n)||^2} \frac{\sigma_{\bar{e}}^2}{\sigma_{\bar{e}}^2 + \gamma \left(\frac{E\{||\mathbf{h}_\Delta(n)||^2\}}{L^2} \frac{\sigma_x^2}{\sigma_x^2} \right) \sigma_v^2}. \tag{73}$$

That is, (73) is also capable of performing the error enhancement by itself. This ensures that the Wiener step-size control is carried out properly when the adaptive filter has not reached sufficient convergence (*i.e.*, for large $E\{||\mathbf{h}_\Delta||^2\}$) or when the mixing system is weakly excited (*i.e.*, for small σ_x^2/σ_v^2). The error enhancement procedure by itself would have difficulty adjusting quickly since it lacks any direct information about the reference signal. In this sense, (73) also performs voice activity detection (VAD) on the reference speech signal to ensure stable adaptation during such an ill-conditioned situation. The

effectiveness of a coring-type nonlinearity that exhibits VAD-like behavior and stabilizes the adaptation of the NLMS algorithm at low signal level was demonstrated in [128], which is consistent with the behavior of (72). Similar behavior should be expected from the oft-used regularized normalization factor $(||\mathbf{x}||^2+\delta)^{-1}$ in (45), which can be used also to perform the Wiener-like error enhancement as long as the parameter δ is controlled accordingly as a function of the noise statistics (*e.g.*, a procedure for adapting δ is provided in [74]).

3.2.2.4 Compressive Nonlinearity for Impulsive Noise

The use of a compressive nonlinearity to limit the onset of double talk essentially suppresses the outliers, *i.e.*, signals corrupted by the impulsive noise, to get back the signal of interest. The strategy is very similar in its functionality to the error enhancement approach. It was shown in [38] that a compressive form of the error nonlinearity can be derived for a bursty noise through the robust statistics theory [54] to improve the performance of the least-squares-type adaptive filters:

$$f_{\text{robust}}(e) = \text{sign}(e)\ \psi\left(\frac{|e|}{s}\right) s, \tag{74}$$

$$\psi\left(\frac{|e|}{s}\right) = \min\left\{\frac{|e|}{s}, k_0\right\}, \tag{75}$$

for $k_0 > 0$, where the scaling parameter s is updated in time as

$$s(n+1) = \lambda_s s(n) + \frac{1-\lambda_s}{\beta}\ \psi\left(\frac{|e(n)|}{s(n)}\right) s(n) \tag{76}$$

for $0 < \lambda_s < 1$ and $\beta > 0$. (74) appears exactly the same in form as (62) in Figure 15(b), which is also able to limit the effect of v generated by a heavy-tailed PDF. The main difference is that (61) and (62) hold the output to within the $\pm\sigma_{\bar{e}}^2/\alpha_v$ range, which allows for an adaptive adjustment of the threshold according to the SNR, while (74) limits the output range to roughly $\pm k_0 s$, where (76) indicates that the adaptive scaling term s represents a low-passed, noise-robust estimate of $E\{|\bar{e}|\}$. Hence (61) and (62) should be able to track the gradual changes in the noise condition better than (74), whereas (74) is more effective than (61) or (62) in limiting the large fluctuations but may be too restrictive such that it leads to slower convergence by the LMS algorithm than necessary during online adaptation.

A compressive nonlinearity that is more general than (74) and goes to zero beyond the threshold can also be derived from the robust statistics theory [146], which is equivalent in effect to freezing the filter adaptation when the residual echo is greater in magnitude than some threshold as a function of the reference signal, *e.g.*, the Geigel DTD algorithm [29]. By following the same reasoning for the origin of the "coring" nonlinearity as discussed in [128], such a "super-compressive" nonlinearity may be obtained when $p_{\bar{e}}$ is Gaussian and p_v is more super-Gaussian than Laplacian.

3.2.3 Analysis of ERN

3.2.3.1 Convergence Behavior with ERN

The convergence behavior in terms of the system distance for the LMS algorithm with the ERN is described approximately by [28]

$$E\{\mathbf{h}_\Delta(n+1)\} \approx (\mathbf{I} - \mu E\{\phi_e'(e(n))\}\mathbf{R}_\mathbf{x})E\{\mathbf{h}_\Delta(n)\}, \tag{77}$$

where $\mathbf{I}$ is the $L\times L$ identity matrix and $\mathbf{R}_\mathbf{x} = E\{\mathbf{x}\mathbf{x}^T\}$ is the $L\times L$ auto-correlation matrix of the reference signal. Then the convergence is guaranteed for

$$0 < \mu < \frac{2}{\lambda_{\max} E\{\phi_e'(e(n))\}}, \tag{78}$$

where $\lambda_{\max}$ is the maximum eigenvalue of $\mathbf{R}_\mathbf{x}$. Refer to [28] for detailed convergence analysis and more precise form of the optimal error nonlinearity, which becomes the same as (70) near convergence for the Gaussian reference signal.

3.2.3.2 Expectation-Maximization Procedure

A recursive and alternating application of the LMS algorithm and the ERN may also be interpreted as following the expectation-maximization (EM) procedure [81]. In the EM algorithm, the maximization step (M-step) is performed by maximizing the joint likelihood of the known signal (*i.e.*, $d(n)$) and the "hidden" signal (*i.e.*, $\bar{d}(n)$), given the old estimate of the "latent" system parameter (*i.e.*, $\mathbf{h}(n)$) to update the parameter estimate (*i.e.*, $\mathbf{w}(n)$), whereas the expectation step (E-step) is performed by determining the conditional expectation of the hidden signal given the observed signal and the new parameter estimate. The

M-step is roughly implemented through the LMS algorithm, where it is well known that the maximum-likelihood (ML) solution is exactly equal to the least-squares (LS) solution when the estimation error is zero-mean Gaussian distributed [56] and that the LMS algorithm stochastically approximates the LS solution [47]. The E-step is precisely what is performed through the Bayesian estimation procedure from which the error enhancement nonlinearities are derived. For example, performing the BSS strictly through the EM algorithm is possible [84], as ICA-based algorithms can also be developed in the ML framework [70].

3.2.3.3 Source Statistics and Overall Form of ERN

Many alternative approaches, statistical or heuristic, may be taken to arrive at the ERN with similar form and functionality. For example, ad-hoc nonlinearities such as the compressed center-clipper and the coring nonlinearity, useful in other signal enhancement scenarios, can be related to the MMSE and the MAP nonlinearities [128]. While the general form of the ERN is dictated by the source statistics, the amount of noise suppression depends directly on the SNR. The major obstacle, however, is that even a typical real-life acoustic mixing environment usually precludes the precise measurement of individual signal statistics. Nonetheless, we show in the remainder of this chapter that rigorous specification of the ERN and estimation of the SNR are not necessary for a sufficient online echo cancellation performance as long as the overall form of the ERN is correctly chosen, given the approximate *a priori* knowledge of the signal distribution, and the entire AEC system is updated via the BIA procedure, which takes advantage of EM-like "dual" combination of adaptation and error enhancement, to take full advantage of the inherent adaptability of the LMS algorithm.

3.2.3.4 Non-Gaussianity of Filter Estimation Error

The Gaussianity of $\bar{e}$ is usually assumed for the LMS algorithm when the filter length is long enough by the argument of the central limit theorem [2]. The error enhancement technique may be interpreted as a generalization of the adaptive step-size method for any probability distribution of $\bar{e}$ or v. That is, the step-size should be adjusted nonlinearly as a function of the signal level for non-Gaussian signals even when their statistics remain stationary. The

ERN technique enables the incorporation of the statistical source information for linear MSE-based adaptive filtering. In fact, the ERN obtained from (60) or (64) suppresses the noise signal better than the Wiener enhancement rule of (67) does when either $\bar{e}$ or v is non-Gaussian distributed (see Appendix B), the implication of which is significant since $\bar{e}$ may not be Gaussian-like during early stages of adaptation if the filter length is relatively short. In any case, most of the signals encountered in real life are not distributed as Gaussian, *e.g.*, the speech signal distribution is widely regarded to be super-Gaussian in either the time or the frequency domain [41]. This leads naturally to the role of ICA as discussed in the next section.

3.2.4 Connection between LMS, ICA, and SBSS

3.2.4.1 Noise Robustness via Independent Component Analysis

A single-channel AEC setup shown in Figure 17(a) can be viewed as a special case of the source separation problem for the recovery of the near-end signals when some of the source signals are partially known, *i.e.*, the far-end (reference) signal. By following the source separation convention, the mixing system in Figure 17(a) can be modeled linearly as

$$\begin{bmatrix} d(n) \\ \mathbf{x}(n) \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{h}(n)^T \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} v(n) \\ \mathbf{x}(n) \end{bmatrix} \tag{79}$$

and the corresponding de-mixing (AEC) system as

$$\begin{bmatrix} \tilde{e}(n) \\ \mathbf{x}(n) \end{bmatrix} = \begin{bmatrix} a(n) & \tilde{\mathbf{w}}(n)^T \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} d(n) \\ \mathbf{x}(n) \end{bmatrix}, \tag{80}$$

where $\mathbf{0}$ is the zero vector of length L and $\tilde{\mathbf{w}}$ is the scaled impulse response estimate vector (the scaling process is discussed in more detail below). Then the natural gradient (NG) algorithm that maximizes the independence between $\tilde{e}$ and $\mathbf{x}$ is given by [144]

$$\tilde{\mathbf{w}}(n+1) = \tilde{\mathbf{w}}(n) + \mu_1[(1 - \phi_{\tilde{e}}(\tilde{e}(n))\tilde{e}(n))\tilde{\mathbf{w}}(n) - \phi_{\tilde{e}}(\tilde{e}(n))\mathbf{x}(n)], \tag{81}$$

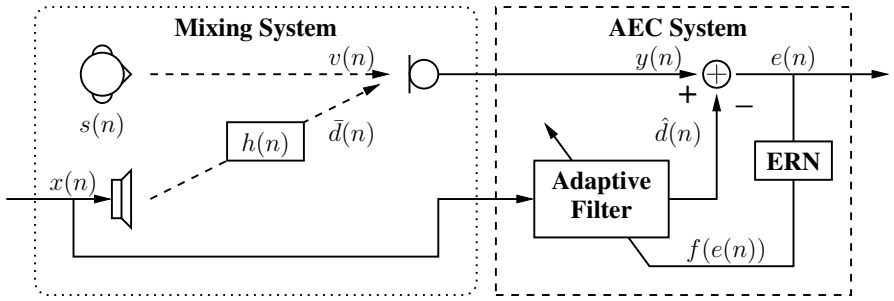
$$a(n+1) = a(n) + \mu_2[1 - \phi_{\tilde{e}}(\tilde{e}(n))\tilde{e}(n)]a(n), \tag{82}$$

for some adaptation step-sizes μ_1 and μ_2. The usual MSE-based system identification scenario is obtained when $a = 1$ and $\tilde{\mathbf{w}} = -\mathbf{w}$ so that $\tilde{e} = d - \mathbf{w}^T\mathbf{x} = e$, where the NG

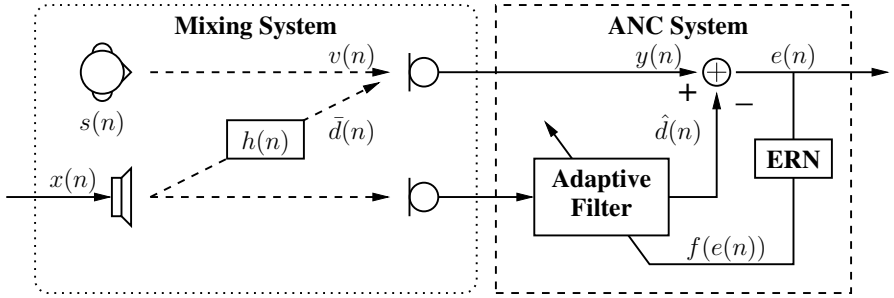
algorithm simplifies to

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu\phi_e(e(n))\mathbf{x}(n), \tag{83}$$

which can be interpreted as the ICA-based LMS algorithm.



(a) Single-channel AEC with the signal $v(n)$ from a local source $s(n)$ added to the desired response $\bar{d}(n)$ (acoustic echo) from a loudspeaker signal $x(n)$ to be canceled.



(b) Single-channel ANC with the signal $v(n)$ form a local source $s(n)$ added to the desired response $\bar{d}(n)$ (noise) from a secondary source $x(n)$ to be canceled.

Figure 17: Comparison between AEC and adaptive noise cancellation (ANC). In both cases, the application of ERN to the filer error $e(n)$ allows an LMS-based adaptive filter to produce the estimate of the target source $s(n)$ that is independent on average from the signal $x(n)$ to be canceled.

By casting the single-channel AEC problem into the two-channel source separation framework with the loudspeaker and the microphone channels as two separate inputs, the ERN falls out automatically from the NG algorithm by applying ICA to the optimization procedure. For example, the mixing system depicted in Figure 17(a) can be generalized for the adaptive noise cancellation (ANC) problem examined in [96] as illustrated in Figure 17(b), which is simply a two-channel BSS problem for recovering one of the source signals. More importantly, the perspective is shifted beyond the identification of the RIR h and focused instead on the estimation of the near-end signal v represented by the output e, *i.e.*, recovery of the original signals by maximizing the independence of the system output signals. The conventional online system identification approaches become ill-conditioned

due to the presence of a near-end source, where historically the adaptive step-size procedure and DTD were developed in retrospect to remedy the stability issue. Then the same assumptions used to guarantee the idealized algorithmic solution in the MSE sense may instead restrict the search for other valid solutions that are, though not necessarily the global optimal, also just as effective. Other significant observations with respect to ICA are as follows.

3.2.4.2 Independence between e and $\mathbf{x}$

The LMS algorithm orthogonalizes (*i.e.*, decorrelates, assuming zero-mean) e and x on average in SOS, *i.e.*, $E\{e\mathbf{x}\} = \mathbf{0}$. On the other hand, the application of ϕ_e to e during the MSE optimization procedure attempts to make e and x independent, which means decorrelation through the second and all high-order statistics (HOS), *i.e.*, $E\{e^k\mathbf{x}\} = \mathbf{0}$ for integer $k \geq 1$. Since the statistical independence implies the second-order decorrelation but not vice versa, the ICA-based LMS algorithm is a generalization of the LMS algorithm for non-Gaussian signals. Gaussian signals are characterized by only up to the SOS, *i.e.*, $\phi_e(e) = e/\sigma_e^2$ when $e \sim Gaussian(0, \sigma_e)$ such that (83) reduces to the LMS algorithm.

3.2.4.3 Connection to Semi-Blind Source Separation

The adaptive scaling factor a of (82) is used to obtain the scaled estimate $\tilde{e} = ad + \tilde{\mathbf{w}}^T\mathbf{x} \approx \tilde{v}$ of the local signal v such that $\tilde{v}$ is "standardized" during the convergence process. Its functionality is similar to that of the scaling factor s of (76) used by the double-talk-robust nonlinearity of (74). The adaptive scaling process makes the NG algorithm robust to large fluctuations in d due to a noise source [144]. Such a mechanism is enforced by the natural gradient learning rule that is scale-invariant, or "covariant" [70], which governs the convergence process of (81) and (82), *i.e.*, $E\{\phi_{\tilde{e}}(\tilde{e})\tilde{e}\} \to 1$ as $n \to \infty$. It has been exploited to perform the single-channel AEC during BSS without a DTD through semi-blind sources separation in the frequency domain in [79]. SBSS is a direct extension of BSS when some partial knowledge of the source signals is already available and is consequently suited for multi-channel AEC in the presence of interferences [89]. Even though the adaptive scaling is fixed for the error enhancement technique, *i.e.*, $a(n) = 1$, the scaling process is an integral

part of the MMSE and the MAP nonlinearities by the virtue of the PDF that standardizes the input signal as a function of the signal statistics. Hence the ICA-based LMS algorithm can be considered as a straightforward alternative to the NG algorithm as long as the SNR can be tracked consistently across time.

3.2.4.4 Intrinsic Enhancement by the Score Function

As motivated above, the ERN derived from the score function is inherently capable of acting as a DTD as ϕ_e is adaptively adjusted as a function of the SNR. Similarly, the NG algorithm of (81) and (82) applies $\phi_{\tilde{e}}$ to $\tilde{e}$, which can be decomposed as

$$\begin{aligned}\tilde{e} &= ad + \tilde{\mathbf{w}}^T\mathbf{x} = (ad + \tilde{\mathbf{h}}^T\mathbf{x}) + (\tilde{\mathbf{w}}^T\mathbf{x} - \tilde{\mathbf{h}}^T\mathbf{x}) \\ &= \tilde{v} + \tilde{\tilde{e}}. \end{aligned} \tag{84}$$

The score function $\phi_{\tilde{e}}$ suppresses the scaled local noise $\tilde{v} = ad + \tilde{\mathbf{h}}^T\mathbf{x}$, where $\tilde{\mathbf{h}}$ is the scaled negative of the RIR vector $\mathbf{h}$, such that a better estimate of the scaled true filter estimation error $\tilde{\tilde{e}} = (\tilde{\mathbf{w}} - \tilde{\mathbf{h}})^T\mathbf{x}$ can be obtained and then be used to update the de-mixing filter coefficients $\tilde{\mathbf{w}}$. Thus the use of a nonlinearity in the NG algorithm may be viewed not only as an essential part of the nonlinear decorrelation process [56, 89] but also as performing the error enhancement to facilitate the filter adaptation. After convergence, (84) must be re-scaled to the original scaling to obtain the local signal estimate [144]. In other words, the error enhancement and the adaptation are both carried out on the scaled signals by the NG algorithm via optimization through ICA, whereas the adaptation is performed entirely in the original scale after the error enhancement by the ICA-based LMS algorithm.

3.2.5 Block-Iterative Adaptation (BIA)

The HOS-based adaptive algorithms are normally suited for batch-wise, offline adaptation such that a mis-specification in the signal statistics, or PDF in general, does not diminish the effectiveness of ICA [56]. There is a trade-off between the offline learning and the online learning, where the former may use as many adaptive iterations as allowed in a given time to further improve the solution, whereas the latter offers the tracking capability in a non-stationary environment. The performance of an ICA-based online adaptive algorithm

depends on how well an adaptation procedure is modified to retain the advantage of batch learning, *e.g.*, the use of so-called "batch-online" adaptation for SBSS in [89] that shortens the batch size considerably yet accomplishes adequate real-time separation performance.

For the LMS algorithm, a small adaptation step-size corresponds to low MSE at the cost of slow convergence speed [46, 49]. The aim is to exploit such an inherent ability of the LMS algorithm to converge to the optimal solution by instilling just enough noise-robustness control to consistently maintain stability. We demonstrated for the single-channel frequency-domain AEC in [131] that a loss in the convergence rate incurred for a gain in the stability due to the error enhancement can then be compensated via BIA without requiring precise estimation of the signal statistics. The EM-like dual re-estimation procedure involving the filter adaptation and the error enhancement permits the refinement of the roughly estimated signal statistics at each iteration until sufficient convergence is obtained. This is different from the conventional data-reuse or batch-wise adaptation philosophy that simply applies the adaptive algorithm repeatedly on a same set of data. Furthermore, it was shown clearly in [8] that the data-reuse NLMS (DR-NLMS) algorithm is similar in form and convergence behavior to the affine projection algorithm (APA). The FBLMS algorithm with BIA requires far less computation than the DR-NLMS algorithm or the APA due to the block processing and the overlap-and-save (or -add) filtering procedure while providing noise robustness during the APA-like adaptation through the error enhancement procedure. We observed experimentally in [131] that four iterations are sufficient for the FBLMS algorithm to maximize the BIA performance, which is analogous to the generally accepted practice of using the fourth-order APA.

The error enhancement technique has also been applied to the traditional multi-channel AEC [132] and combined with a RES [142, 143] with excellent results. The successes are a testament to the system approach to signal enhancement that culminates in the robustness against mis-estimation of statistics. Most of the existing ICA-based BSS procedures already utilize *a priori* knowledge of the source statistics and batch-wise adaptation along with the natural gradient algorithm to provide sufficient source separation. The same framework

underlies the error enhancement paradigm for AEC, realized through the system combination of the LMS algorithm, the ERN, the adaptive regularization procedure (which can be consider an integral part of the error enhancement component), and the BIA procedure. In such a framework, where the system components are designed to interact mutually with each other, prior source information, which dictates the overall characteristics of the ERN, and not necessarily the precisely estimated signal statics becomes essential for robust AEC performance.

3.3 Experimental Evaluation

The standard usage of VAD to estimate the noise variance during silence for the error enhancement has already been tested in [128, 129, 130]. Other advanced techniques, *e.g.*, [124], may be employed for improved tracking performance. However, the conventional single-channel techniques tend to be heuristic in nature and depend strongly on the arbitrarily chosen thresholds and smoothing constants. In order to clearly demonstrate the behaviors of different combinations of the ERN, the given SNR (averaged over time for entire data) is used throughout the time-domain simulation. The main purpose is to show that the sensitivity to the choice of fixed parameters is reduced by the system approach and that the precisely estimated signal statistics are not necessary through the BIA procedure, which can be carried out efficiently in the frequency domain. Thus the full practical potential of the error enhancement technique is illustrated by the frequency-domain simulation.

A simulated acoustic echo, generated from a RIR with the reverberation time of about 250 ms measured at 8 kHz sampling rate, was used. The RIR was truncated prior to mixing to match its length to that of the adaptive filter, chosen here to be 64 ms, and force zero filter-length mismatch error (again for the purpose of illustrating the behavior of the ERNs more clearly), and it was scaled to produce the echo return loss (ERL, *i.e.*, attenuation) of 10 dB. 16 kHz, 16-bit PCM speech from the TIMIT database downsampled to 8 kHz were used.

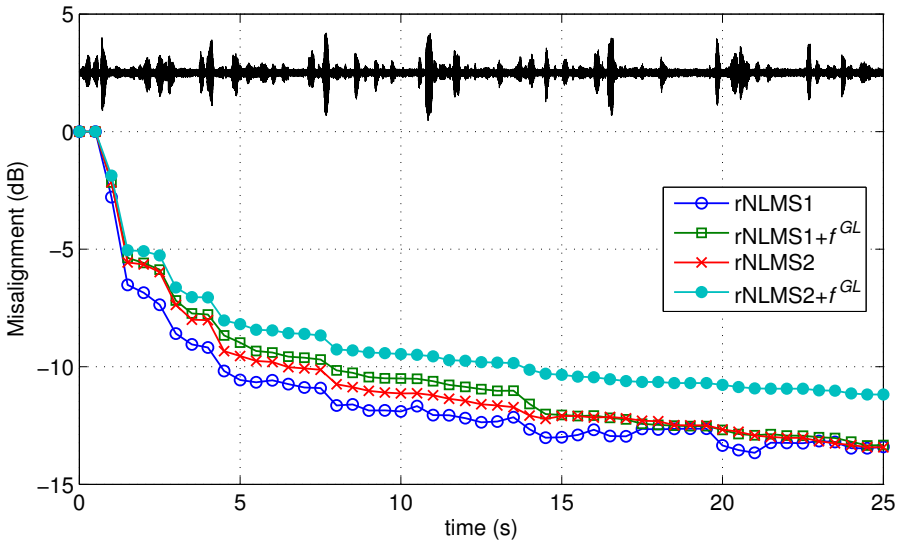
Three main scenarios were examined. First, an ambient noise comprised of white Gaussian noise (WGN) was added to the acoustic echo during the time-domain AEC to simulate

a linear distortion. Second, encoding and decoding using the GSM AMR speech codec [58] were applied to the acoustic echo during the time-domain AEC to simulate a nonlinear distortion. Finally, a continuous near-end speech was added along with the WGN during the frequency-domain AEC to simulate a very noisy AEC condition that would normally prohibit the use of conventional single-channel noise estimation techniques.

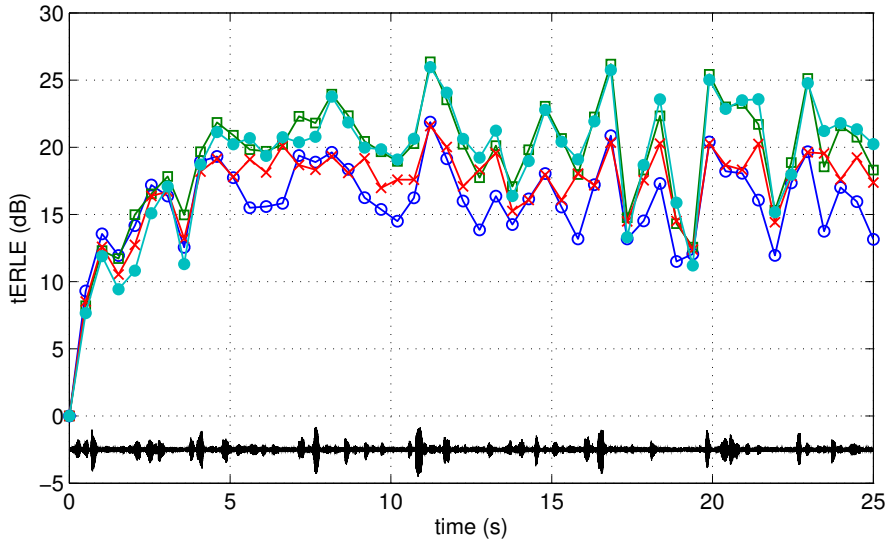
3.3.1 Time-Domain AEC with Ambient Noise

Figure 18 shows the results from a simulated acoustic echo with additive WGN at 10 dB SNR and using a compressive ERN ($f_{\text{MMSE}}^{\text{GL}}$) with known noise energy. The regularization parameters, δ for the first regularized NLMS algorithm in (45) (rNLMS1) and γ for the second regularized NLMS algorithm in (72) (rNLMS2), were adjusted to give the highest steady-state tERLE when the step-size of $\mu = 1$ is used. We arrive at several observations from the figure. First, rNLMS2 performs better than rNLMS1 when the SNR is low, *i.e.*, at low signal level, which is consistent with the intended design of rNLMS2. Second, combining $f_{\text{MMSE}}^{\text{GL}}$ with rNLMS1 or rNLMS2 allows as much as 5 dB improvement in the tERLE. Third, the tERLE plot clearly indicates that echo cancellation beyond the noise level (10 dB *a priori* SNR) is indeed possible through a combination of the error enhancement (adaptive step-size) and regularization procedures. Finally, higher steady-state tERLE is obtained with the inclusion of the ERN technique even though the misalignment is higher when compared to the performances of rNLMS1 or rNLMS2 without the ERN. Thus lower misalignment, or better system identification, does not always translate to higher echo cancellation performance in a noisy situation.

Figure 19 illustrates the stabilizing effect of the ERN on rNLMS2 for the mis-specification of γ, which also implies to a large extent the robustness against the mis-estimation of the signal statistics. Although rNLMS2 is capable by itself of handling the presence of an ambient noise, $f_{\text{MMSE}}^{\text{GL}}$ enables rNLMS2 to achieve even higher tERLE (averaged over the last 10 seconds of simulation to ensure sufficient convergence) by providing increased noise robustness when the regularization parameter is under-specified. The adaptive regularization procedure of rNLMS2 applies a Wiener-like step-size correction, which dynamically



(a) Misalignment.



(b) True ERLE.

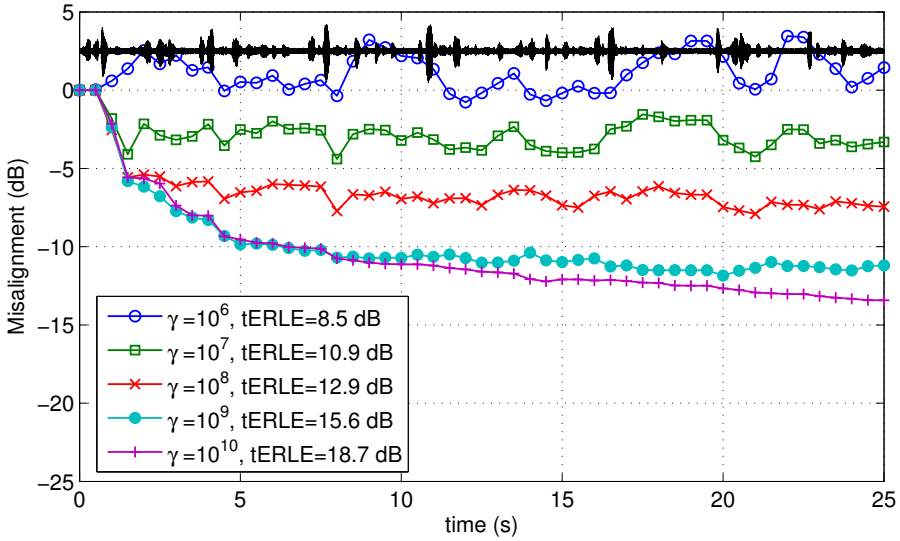
Figure 18: From the regularized NLMS algorithms (rNLMS1, rNLMS2) with the ERN ($f^{\mathrm{GL}}_{\mathrm{MMSE}}$).

emulates a coring nonlinearity [128] and enhances the filter estimation error at low signal level, whereas a combination of rNLMS2 and $f_{\text{MMSE}}^{\text{GL}}$ ensures the proper enhancement for a wide range of time-varying SNR resulting from the non-stationary speech signal. The same overall behavior is observed for rNLMS1 when δ is varied.

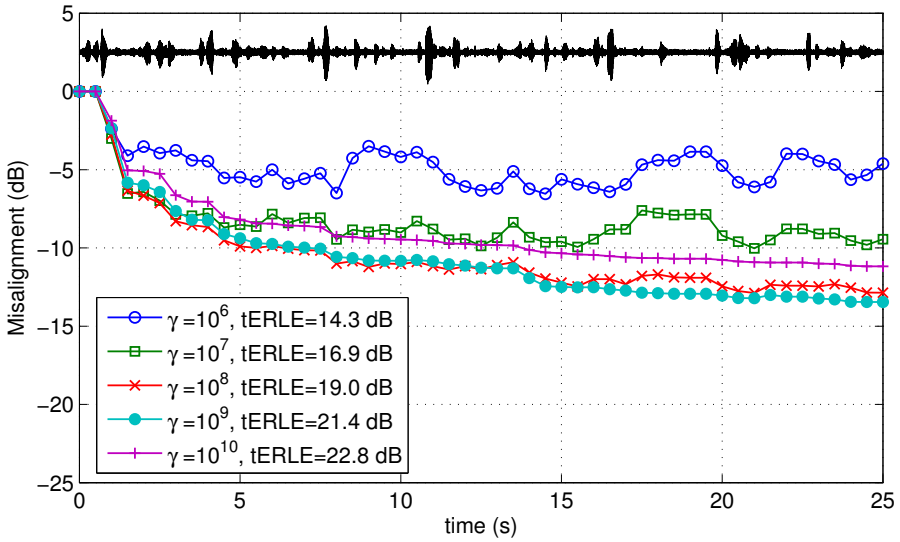
3.3.2 Time-Domain AEC with Speech Coding Distortion

By modeling the nonlinear speech coding distortion as an additive noise, *i.e.*, $v(n) = \bar{y}(n) - \hat{y}(n)$ for the original speech $\bar{y}$ and a distorted speech $\hat{y}$, the signal-to-distortion ratio (SDR) can be defined in the same fashion as the SNR. Then the error enhancement procedure may also be applied to the network-based AEC as illustrated in Figure 20(a) (communication and processing delays between the encoder and the decoder are ignored). The amount of distortion is signal-level dependent and may be pre-determined as indicated in Figure 20(b), where the GSM AMR coding distortion (measured in terms of the SDR and averaged over 20 female speech utterances) is plotted against the input speech magnitude (measured in terms of the signal loss in dB) for various coding rates.

Figure 21 is obtained from a simulated acoustic echo encoded and decoded by GSM AMR at 12.2 kbps bit-rate. The regularization parameter for rNLMS2 was adjusted to give the highest steady-state ERLE when $\mu = 1$ is used. Instead of attempting to accurately model the amount of distortion in terms of the input signal magnitude as in [128], a fixed distortion estimate was used this time throughout the simulation just as a fixed noise power estimate was used for the previous ambient noise case. Compared to the increase in the average ERLE by 3 dB as reported in [128], more than 7 dB in the improvement is obtained this time for NLMS+$f_{\text{MMSE}}^{\text{GG}}$ over NLMS alone. It suggests that one must be careful how the step-size is adjusted when dealing with nonlinear distortions, in which case it is better to let the LMS algorithm converge naturally without exerting more secondary control than necessary. rNLMS2 by itself provides the same overall performance as NLMS+$f_{\text{MMSE}}^{\text{GG}}$, which is as expected since rNLMS2 already possesses the Wiener-like error enhancement capability. As such, the application of $f_{\text{MMSE}}^{\text{GG}}$ to rNLMS2 does not provide any substantial benefit.

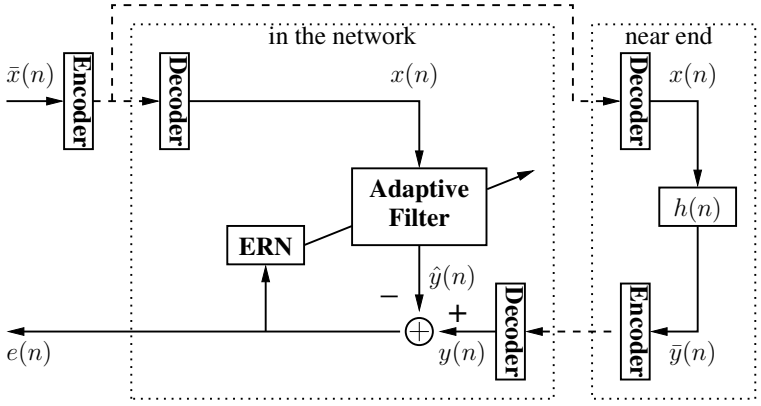


(a) Without $f^{\mathrm{GL}}_{\mathrm{MMSE}}$.

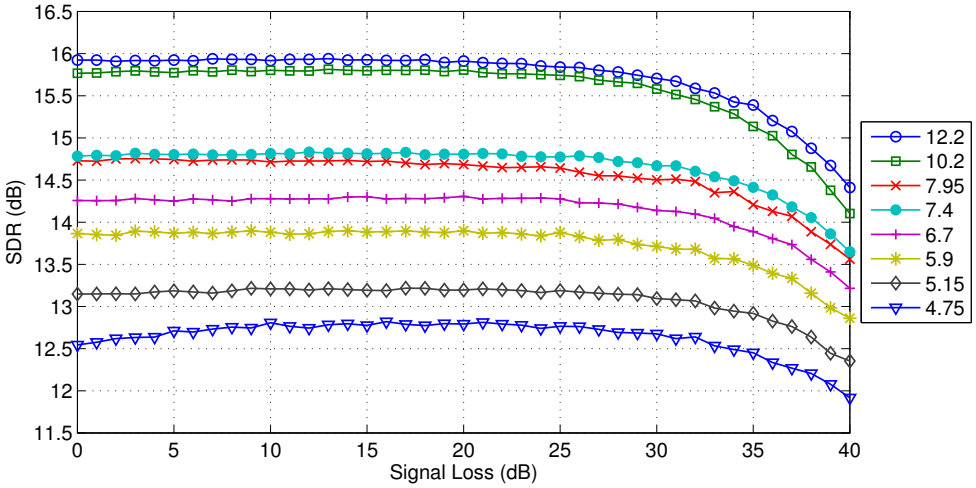


(b) With $f^{\mathrm{GL}}_{\mathrm{MMSE}}$.

Figure 19: From the regularized NLMS algorithm (rNLMS2) with the ERN ($f^{\mathrm{GL}}_{\mathrm{MMSE}}$). Corresponding average tERLE is provided in the legend.

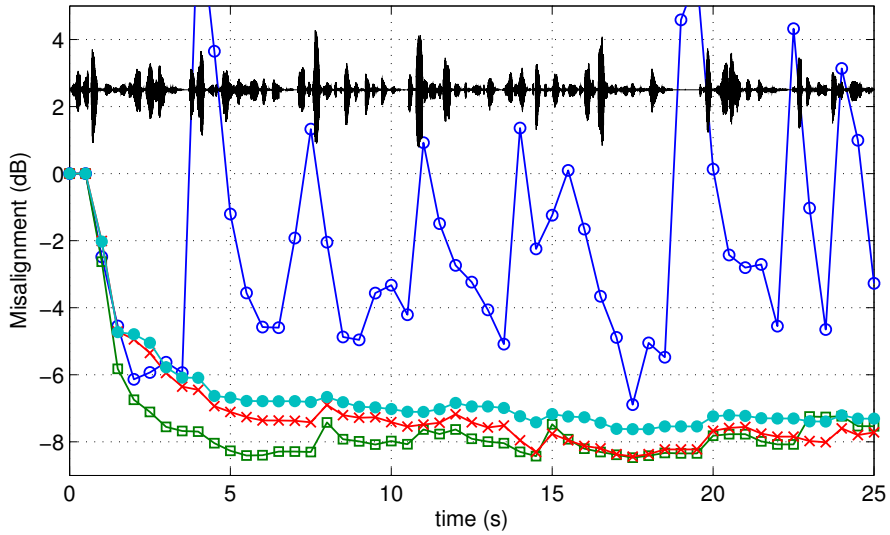


(a) Model for network-based AEC with the ERN.

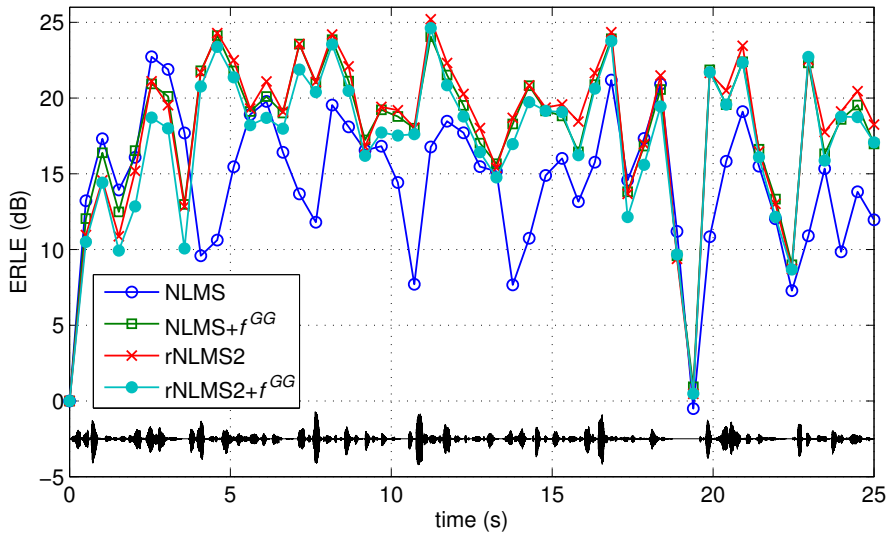


(b) Signal-to-distortion ratio (SDR) versus signal loss.

Figure 20: Network-based AEC with speech coding distortion for various bit-rates (kbps). Quantization noise starts to take over beyond the signal loss of 25 dB.



(a) Misalignment.



(b) ERLE.

Figure 21: From the NLMS algorithms (NLMS, rNLMS2) with the ERN ($f^{\mathrm{GG}}_{\mathrm{MMSE}}$). GSM AMR speech encoding and decoding at 12.2 kbps bit-rate were applied to the acoustic echo.

We have observed that using other types of ERNs with NLMS does not improve the ERLE much more than $f_{\text{MMSE}}^{\text{GG}}$ for the speech coding distortion. This is also expected since GSM AMR by design imparts perceptually weighted distortion in the frequency domain, which results in the fixed SDR on average except at low signal level when the distortion is mainly due to the quantization noise as indicated by Figure 20(b). In such a case, $f_{\text{MMSE}}^{\text{GG}}$ that dynamically mimics a coring nonlinearity well and does not overly alter the input signal should serve better than either a subtractive or a compressive ERN. This is a good example of how the *a priori* knowledge of source characteristics and the understanding of the actual physics involved influence the proper integration of system components for delivering the best performance possible.

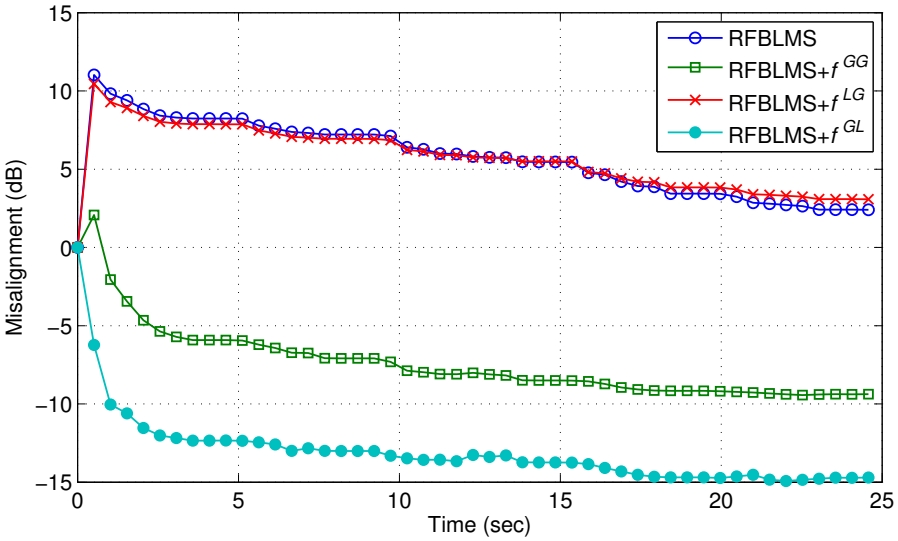
3.3.3 Frequency-Domain AEC with Ambient Noise and Double-Talk

Figure 22 shows the results from using all the three types of ERNs ($f_{\text{MMSE}}^{\text{GG}}$, $f_{\text{MMSE}}^{\text{LG}}$, $f_{\text{MMSE}}^{\text{GL}}$) with the regularized FBLMS (RFBLMS) [129] on a simulated acoustic echo, where silence (WGN at 90 dB SNR) of 1 second was inserted between the reference speech utterances before the near-end mixing. Both the WGN at 10 dB SNR and a continuous local speech at 0 dB SNR were added to the acoustic echo. The same simplified statistics estimation strategy in [129] was utilized, *i.e.*, $v \approx e$ and $\xi = 1$ (in such a case, $f_{\text{MMSE}}^{\text{GL}}$ and $f_{\text{MAP}}^{\text{GL}}$ become much like the compressive nonlinearity of (74) by limiting e to be roughly within $\pm\sigma_{\bar{e}}$, *e.g.*, for $f_{\text{MAP}}^{\text{GL}}$ the threshold becomes $t = \xi\alpha_v \approx \sigma_e$ for the SNR $\xi = \sigma_{\bar{e}}^2/\alpha_v^2$), as well as $\gamma = 1$ for the regularization procedure. The ERN was applied only to the input magnitude while the phase was left unmodified. Relatively large adaptation step-size $\alpha = 0.02$ and smoothing constant $\beta = 0.998$ were used for RFBLMS such that only four BIAs (*i.e.*, filter→ERN→adaptation) per batch of data were necessary for RFBLMS+$f_{\text{MMSE}}^{\text{GL}}$ to reach the maximum steady-state tERLE. The tERLE plot shows that RFBLMS+$f_{\text{MMSE}}^{\text{GL}}$ provides the best result during the entire simulation. The misalignment plot also exhibits the stability of RFBLMS+$f_{\text{MMSE}}^{\text{GL}}$, where the large α causes instability at the start of adaptation from which the other combinations are unable to recover. The noise robustness of RFBLMS+$f_{\text{MMSE}}^{\text{GL}}$ is also apparent as high tERLE is maintained even during the silences in

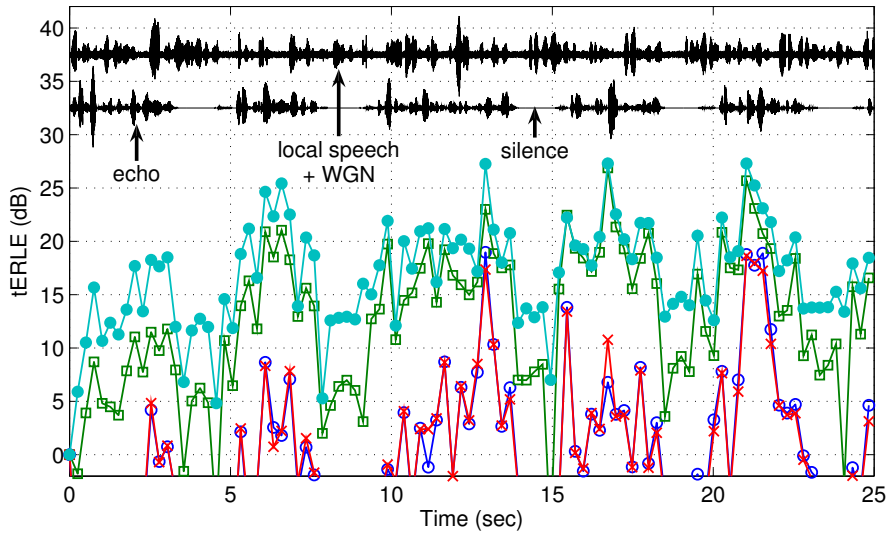
the acoustic echo signal.

Figure 23 illustrates the APA-like behavior of RFBLMS with $f_{\mathrm{MMSE}}^{\mathrm{GL}}$ for various block iterations *iter* and step-size α. A significant recovery of the convergence speed, especially at the initial adaptation stage, is possible through BIA. The AEC system is able to recover after a sudden change in the RIR at 10 seconds. We note that although the acoustic echo is still audible at around 20 dB tERLE, the use of a RES and the masking effect during double talk must be taken into account in practice [142]. Also, this simulation example does not imply that a DTD should not be used at all; rather, it should be incorporated in such a manner to help the overall system performance, *e.g.*, the step-size can be decreased by a factor of half during double talk for extra stability [132]. The current practice of freezing the adaptation entirely, depending on a particular choice of the DTD threshold, may be more disruptive than helpful due to, for example, the false detection or when the RIR changes during double talk.

The double-talk situation can create large variations in the filter estimation error, especially during the frequency-domain AEC when a speech signal is very sparsely represented across frequency. $f_{\mathrm{MMSE}}^{\mathrm{GL}}$ performs the best out of the three types of ERNs since it is able to suppress the large outliers in the observed estimation error just as with the traditional technique of using a compressive-type nonlinearity to counter the double-talk leakage. As the BIA procedure enables the recovery of the convergence rate lost due to the scaling down of the step-size, such a mechanism becomes crucial during double talk that requires a more stringent step-size control than for the ambient noise. None of the previously mentioned frequency-domain AEC algorithms designed specifically to work with the double-talk situation [126, 114] were able to take advantage of batch-wise adaptability afforded by the error enhancement procedure. Replacing $f_{\mathrm{MMSE}}^{\mathrm{GL}}$ with $f_{\mathrm{MAP}}^{\mathrm{GL}}$, which is much simpler to implement numerically, results in nearly identical behavior as shown in Figure 24 for the same step-size (slight differences in the performance may be adjusted by varying the step-size). Thus the proper specification of the overall form of the ERN, dictated by the *a priori* source information, is practically more relevant than that of the exact mathematical form.

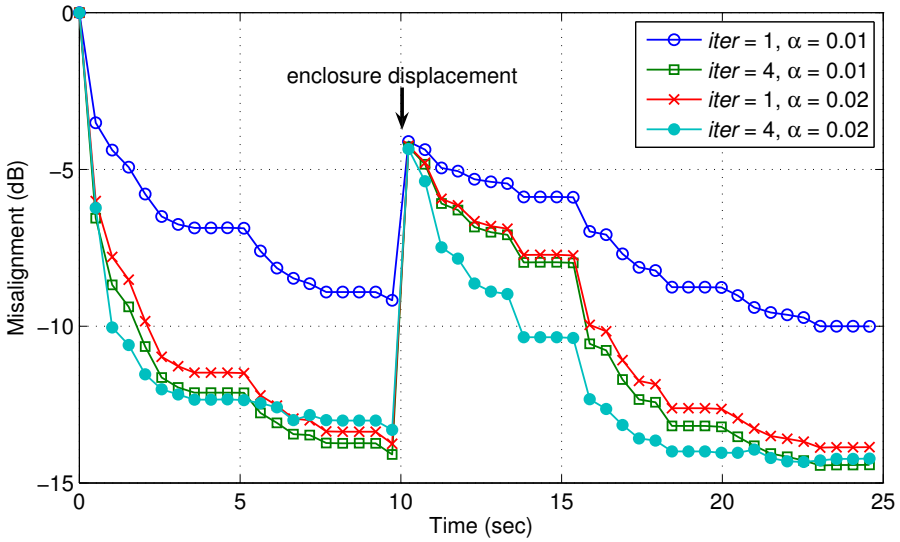


(a) Misalignment.

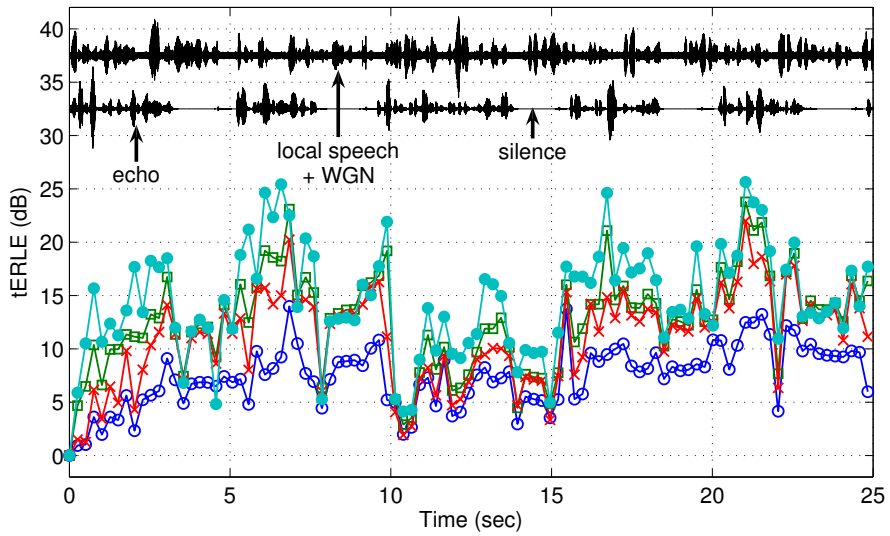


(b) True ERLE.

Figure 22: From the regularized FBLMS algorithm (RFBLMS) with the ERN ($f^{\mathrm{GG}}_{\mathrm{MMSE}}$, $f^{\mathrm{LG}}_{\mathrm{MMSE}}$, $f^{\mathrm{GL}}_{\mathrm{MMSE}}$).

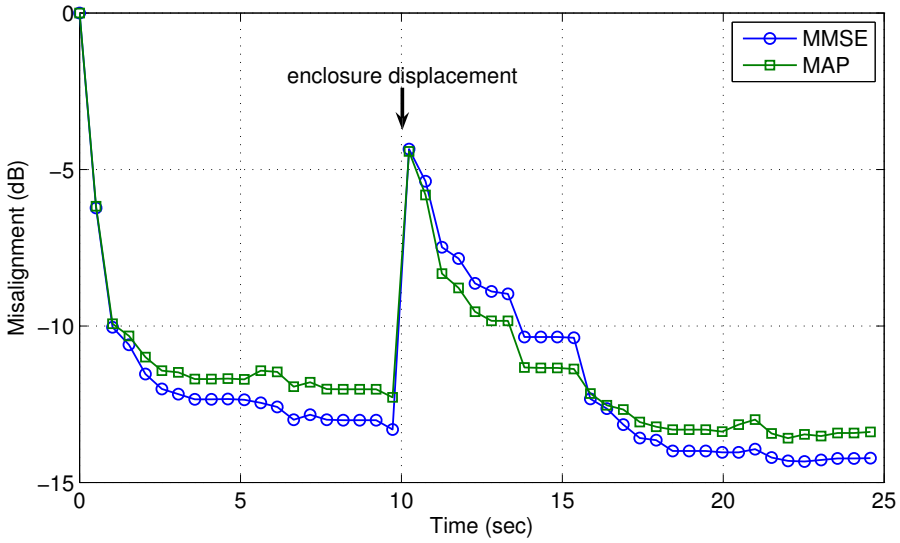


(a) Misalignment.

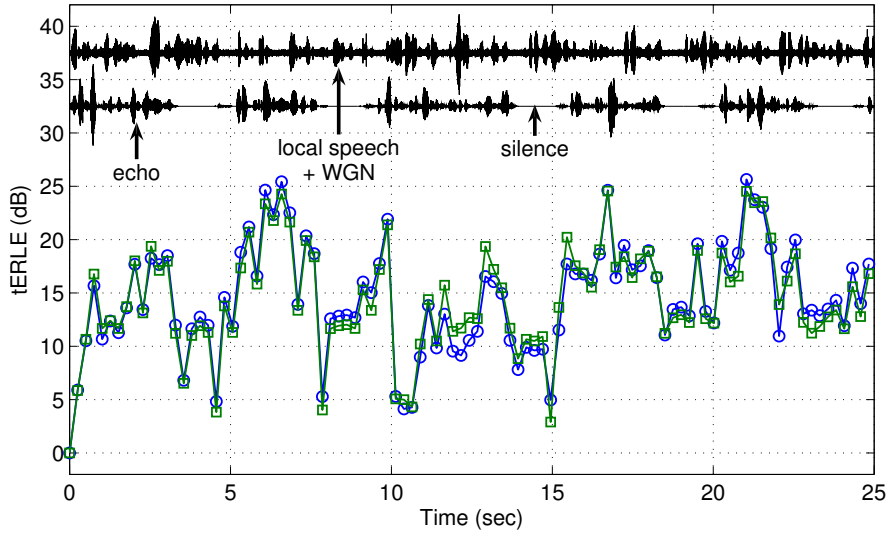


(b) True ERLE.

Figure 23: From RFBLMS with $f_{\text{MMSE}}^{\text{GL}}$ for various block iterations *iter* and step-size α. The RIR changes suddenly at 10 seconds due to the displacement of the loudspeaker-microphone enclosure.



(a) Misalignment.



(b) True ERLE.

Figure 24: From RFBLMS with $f^{\mathrm{GL}}_{\mathrm{MMSE}}$ or $f^{\mathrm{GL}}_{\mathrm{MAP}}$.

CHAPTER IV

SYSTEM PERSPECTIVE OF DECORRELATION FOR AEC

As with any other least-square-based adaptive filtering, the least mean square (LMS) algorithm suffers from two main problems. One is that the convergence speed is decreased greatly for highly correlated reference signal x, *e.g.*, speech, thereby making the normal equation

$$\mathbf{R}_{xx}\mathbf{w} = \mathbf{r}_{xy} \tag{85}$$

ill-conditioned for solving numerically, where $\mathbf{R}_{xx} = E\{\mathbf{x}\mathbf{x}^T\}$ is the auto-correlation matrix, $\mathbf{x}$ and $\mathbf{w}$ are the reference and the filter coefficients vectors, respectively, $\mathbf{r}_{xy} = E\{\mathbf{x}y\}$ is the cross-correlation vector, y is the microphone signal, and $\{\cdot\}^T$ is the transpose operator. Another problem is the adaptive algorithm's sensitivity to the distortion on y, *i.e.*, $y = d + v$ for linear distortion where v can be the near-end speech or the background noise, that consequently corrupts the error e needed for ideal filter coefficients updating, *i.e.*, $e = y - \hat{d} = b + v$ where $b = d - \hat{d}$ is the "true" (*i.e.*, noise-free) error. That is, the LMS algorithm by itself has difficulty converging to the optimal solution in the presence of local noise (*e.g.*, double talk) since the noise directly perturbs the single-sample, "noisy" estimate of the MSE gradient (*i.e.*, the gradient, or estimation, noise [49]). Therefore, the ill-conditioned $\mathbf{R}_{xx}$, the near-end noise v, and the gradient noise all together conspire to degrade the AEC performance in a disruptive and complex acoustic mixing system.

There are other issues associated with the LMS algorithm in the real-world AEC scenarios. Namely, the "non-uniqueness" problem arises due to *inter-channel correlation* during multi-channel AEC (MCAEC) that further retards the convergence speed [34]. To improve the echo-path tracking, a decorrelation procedure must be applied to the reference signal before playback and adaptation at the cost of decreased audio quality. A similar problem occurs due to *inter-block correlation* with the multi-delay filter (MDF) [120], which is a partitioned-block approach to the frequency-block LMS (FBLMS) algorithm [49] for reduced

computation and blocking delay. The two related issues can be alleviated by accounting for the inter-correlation structure [19]. This, however, ultimately involves the inversion of the auto-correlation matrix, where a fast, but potentially unstable, algorithm may be utilized for the inversion process, *e.g.*, [18].

In the previous chapter, we showed the systematic relationship in detail between the LMS algorithm and the error recovery nonlinearity (ERN) during residual echo enhancement (REE) for noise-robust AEC performance [131, 133]. The system approach, which places analytically consistent yet global perspective on the problem at hand rather than focusing on individual, idealized algorithm, permits robust performance in real-world scenarios. The noise-robustness issue can be effectively solved by REE that applies the ERN to "enhance" the filter estimation error before updating the filter coefficients. Both the steady-state and the convergence behaviors of the LMS algorithm are improved significantly through REE and multiple recursive filtering and adaptation on a batch of noisy data via block-iterative adaptation (BIA). However, while the REE technique may be readily extended to the traditional single-channel solution to MCAEC, it does not directly address the non-uniqueness issue to reduce the dependency of the Wiener solution on the far-end room response. Some form of decorrelation must still be used with aforementioned trade-offs.

Key criteria for an ideal decorrelation procedure are as follows.

- Retains the original audio quality and image of far-end sources.
- Retains the original excitation characteristics of echo paths.
- Retains the original signal statistics used for adaptive filtering.
- Extendable to large number of channels.
- Requires low computational complexity.

Many conventional techniques, *e.g.*, a nonlinear "half-wave rectifying" processor [34] and comb filtering [11], do not entirely satisfy the first two requirements. They may not likely meet the third condition necessary for optimal steady-state performance by an MSE-based

adaptive filter, and they also tend to be incompatible for the case of more than two audio channels.

We will demonstrate in this chapter that the same system perspective can be extended to the decorrelation procedure for directly assisting the AEC adaptation process. The overall aim here is to achieve decorrelation by integrating the decorrelation procedure not simply as a separate pre-processor applied prior to the LMS algorithm but as a part of the AEC system, capable of controlling both the echo cancellation and tracking performances while introducing the least amount of audio distortion possible.

The rest of this chapter is organized as follows. First in Section 4.1, we present the decorrelation by resampling (DBR) technique that exploits the systematic link between BIA, decorrelation, and resampling. The new technique leads to the development of frequency-domain resampling (FDR), which takes advantage of the computational efficiency of the Fast Fourier Transform (FFT), and sub-band resampling (SBR), which is an extension of FDR that allows selective decorrelation per frequency sub-band as measured by the coherence [20] for controlling the trade-off between signal distortion and decorrelation amount. Second in Section 4.2, we provide extension of the techniques already developed and used for a single-channel robust AEC (R-AEC) system to MCAEC and MDF, where we also discuss the variations in BIA for further improvement in the AEC performance. Finally in Section 4.3, all the proposed DBR techniques are tested for audio quality and AEC. In particular, we evaluate the true echo return loss enhancement (tERLE) and the misalignment through the sub-band decomposition to demonstrate the superiority of our system approach over other individual decorrelation methods.

Our stance is markedly different from the traditional aspects in several ways. First, the DBR and the R-AEC components complement one another such that low coherence, or equivalently low misalignment, over the entire frequency range is not necessary for high tERLE during MCAEC. Conventional wisdom instead favors largest decrease in the coherence as possible over all frequencies, which most likely causes the degradation of the actual cancellation performance itself [132, 137]. Second, the R-AEC component takes advantage of intrinsic adaptability of the LMS algorithm by applying BIA to the FBLMS algorithm

for increased convergence speed rather than simply relying on fast-converging alternatives such as the affine projection algorithm or the recursive least square algorithm. Finally, the R-AEC component, which includes the error enhancement and the adaptive regularization procedures [131, 133], is adapted continuously in noise-robust fashion via REE even during double talk. Such a robustness, in turn, is essentially what makes BIA possible. We illustrate in this chapter the first two points above through the sub-band analysis of the tERLE and the misalignment obtained from MCAEC and MDF.

4.1 Inter-channel Decorrelation by Resampling

4.1.1 Review of the Non-uniqueness Problem

Brief review of the inter-channel correlation problem is as follows.

Let $y_i(n) = \sum_j \sum_k h_{ij}(k)x_j(n-k) = \sum_j \mathbf{h}_{ij}^T \mathbf{x}_j(n)$ be the noise-free recording from i^{th} microphone, where "T" denotes vector transposition, $\mathbf{x}_j(n)$ is the reference vector from j^{th} loudspeaker, $\mathbf{h}_{ij}$ is the time-invariant room response vector, $1 \le i \le P$, $1 \le j \le Q$, $0 \le k \le N-1$, and N is the near-end room impulse response (RIR) length. Assuming $L = N$, a set of filter coefficients corresponding to the echo paths between all Q loudspeakers and i^{th} microphone is given by

$$\{\hat{h}_{ij}(n)\}_i = \underset{w_{ij}}{\operatorname{argmin}} E\left(y_i(n) - \sum_j \sum_k w_{ij}(k)x_j(n-k)\right)^2, \tag{86}$$

which leads to the multi-channel normal equation

$$E\left[\left(y_i(n) - \sum_k w_{ij}(k)x_j(n-k)\right)x_{j'}(n-k')\right] = 0, \tag{87}$$

or represented more compactly as the familiar normal equation

$$\mathbf{R}\mathbf{w}_i = \mathbf{r}_i, \tag{88}$$

where $\mathbf{R} = \{E\left[x_j(n-k)x_{j'}(n-k')\right]\}$ is an $LQ \times LQ$ matrix, $\mathbf{w}_i = \{w_{ij}(k)\}_i$ and $\mathbf{r}_i = \{E\left[y_i(n)x_{j'}(n-k')\right]\}_i$ are $LQ \times 1$ vectors, and $E\{\cdot\}$ is the expectation operator.

The normal equation indicates that even if the uniqueness condition of $L < M$ is met (*i.e.*, $\mathbf{x}_j$ is linearly independent of $\mathbf{x}_{j'}$ for $j \ne j'$ [57]), which is most likely the case in reality, the problem remains ill-conditioned if $E\left[x_j x_{j'}\right] \ne 0$. The convergence behavior of

a stochastic gradient descent algorithm is then assisted by a decorrelation procedure $\phi(\cdot)$ such that $E\left[\phi(x_j)\phi(x_{j'})\right] \approx 0$ for $j \neq j'$. Still, any extra processing, linear or nonlinear, will inevitably change the statistics of non-stationary random processes $x_j(n)$ and $x_{j'}(n)$ and modify the steady-state (or near steady-state) solution, where the effect may be significant for the LMS algorithm that uses a very rough estimate of the gradient. A decorrelation procedure should be designed to minimize such an effect.

4.1.2 Systematic Link between BIA, Decorrelation, and Resampling

The REE technique is represented systematically in Figure 25 that recursively refines the estimations $\hat{b}$ and $\hat{v}$ of the corrupted residual echo $e = b+v$. This built-in dual re-estimation process, carried out via BIA, makes the R-AEC component based on the LMS algorithm less sensitive to mis-specification of the system parameters and mis-estimation of the signal statistics [133]. BIA also permits a natural recovery of the convergence speed lost to the coloration of the reference signals, *e.g.*, due to the non-uniqueness problem [132, 137].

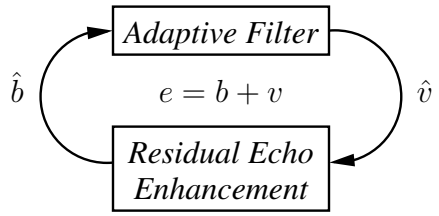


Figure 25: System integration of adaptive filter and REE.

The LMS algorithm iteratively and stochastically solves the normal equation (85). For such a dynamic solution, a mismatch in the sampling rate between the loudspeaker and the microphone channels on the order of few hundred parts per million ($\sim 0.01\%$) is enough to break down the correlation structure of $\mathbf{r}_{xy}$ in (85) for significant decrease in the LMS algorithm's cancellation performance [100]. Conversely, we should be able to induce the same effect for the decorrelation purpose, *i.e.*, to improve the conditioning of $\mathbf{R}_{xx}$, by resampling x frame-wise. Such a close and dynamic relationship between the sampling rate and the LMS algorithm is one major reason for DBR's effectiveness as revealed in [132, 137].

One other crucial system aspect is that BIA enables a natural recovery of the convergence

speed and hence reduces the need for aggressive decorrelation applied directly to x, which subsequently minimizes the audio distortion and also the potential interference with the adaptive cancellation process. We have observed that the DBR and the R-AEC components complement one another such that low coherence, or equivalently low misalignment, over the entire frequency range is not necessary for high cancellation performance during MCAEC. Conventional wisdom instead favors as large decrease in the coherence as possible over all frequencies, which most likely causes the degradation of the actual cancellation performance [132, 137].

4.1.3 Decorrelation by Resampling (DBR)

Let f_s and f_s' be the original and the new sampling rates, respectively. The resampling ratio is defined as

$$R = \frac{f_s'}{f_s} = 1 + r, \tag{89}$$

where the mismatch ratio is defined as $r = f_\Delta / f_s$, $f_\Delta = f_s' - f_s$. Assuming without loss of generality a real-valued $R > 1$ (or $r > 0$), sampling rate expansion gives the identity relationship

$$x(nR^{-1}) \longleftrightarrow X(Z^R), \tag{90}$$

where $x(n)$ and $X(Z)$ are the discrete time sequence of a continuous time signal $x(t)$ and the corresponding Z-transform, respectively, after which the upsampled signal is obtained by lowpass filtering (interpolation) [93]. By equating $x(n-d) = x(nR^{-1})$,

$$d = n\left(\frac{r}{1+r}\right), \tag{91}$$

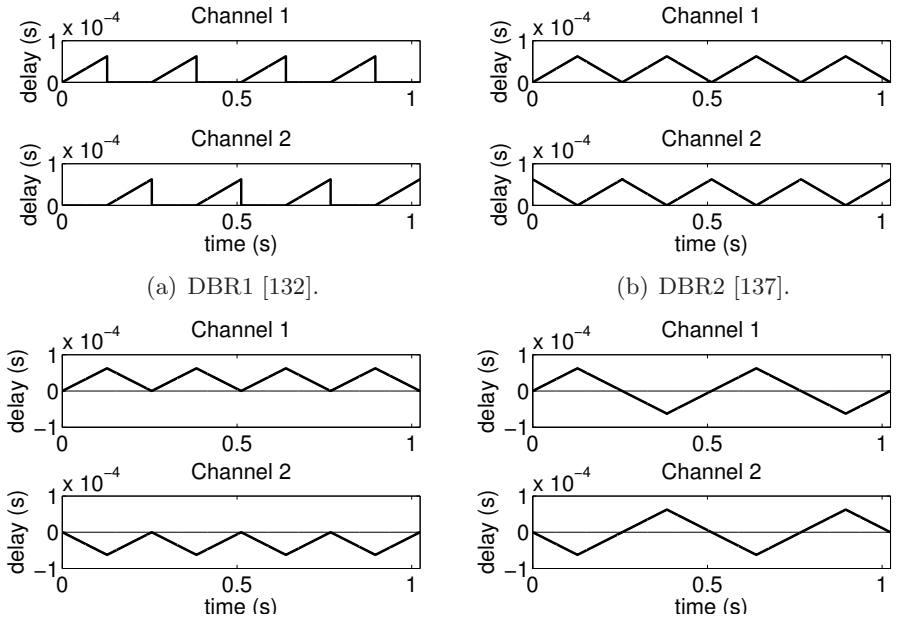
which is the fractional delay of expanded samples with respect to the original samples. Thus after upsampling, (90) implies spectral warping (*i.e.*, frequency-dependent modulation) is applied to the original signal, and (91) means the delay grows progressively in time (*i.e.*, samples gradually accumulate over the original time scale).

A time-varying phase shift in sub-bands was applied as a decorrelation procedure for MCAEC in [51] with larger modulation at higher bands to perceptually hide the signal distortion after synthesis, whereas one-sample delay was inserted periodically across channels

into frames with half delay period per frame and quarter-period shifting during stereophonic AEC (SAEC) in [125]. We propose combining the resampling approach with the alternating projection technique of [110]. Such a combination takes on key features from [51] and [125] as it periodically imparts smoothly increasing modulation (in frequency) and delay (in time) across channels. The main drawback is the computational cost of resampling at the rate $R \simeq 1$ ($r \simeq 0$), which requires very large integer-valued resampling ratios for the ideal upsampling and downsampling scheme. Such a problem can be solved by the resampling-by-interpolation strategy proposed in [100], which drastically reduces the computation time by omitting the downsampling process and reusing a short interpolation filter (sinc function) per block. Therefore, the decorrelation is achieved simply by lowpass filtering in an appropriate manner.

Figure 26 demonstrates several ways to apply DBR for inter-channel decorrelation. The resampling rate R is adjusted arbitrarily here to produce the nearest whole extra sample for the given frame size N. For DBR1, resampling is applied to every other channel and time frame, which then requires a smoothing procedure to ensure continuity across the resampled frames [132]. On the other hand for DBR2, DRB3, and DBR4, resampling is applied simultaneously across all channels by time-reversing a frame in every other channel before resampling at the opposite rate than the previous frame, *i.e.*, $1/R$ versus R, and reversing back afterward. This results in continuously varying delay, thus continuity in the resampled signal. Negative delay in Channel 2 for DBR3 and DBR4 is produced by resampling every other frame at the rate $1/R$, $R > 1$, and by time-reversing and resampling the rest at R and reversing them back afterward. Many other framing procedures are of course possible for smooth delay variation across time.

Figures 27 and 28 show the corresponding inter-channel coherence plots. The coherence decreases with N as R is increased accordingly ($N = \infty$ indicates no DBR). DBR1, DBR2, and DBR3 provide results very similar to each other, where DBR1 gives slightly lower coherence at higher frequencies than others due to the undesirable aliasing distortion caused by decimation, or discontinuity in delay, that occurs across frames [125]. DBR4 is able to lower the coherence more than others since it applies as much continuous delay variation



(a) DBR1 [132].

(b) DBR2 [137].

(c) DBR3.

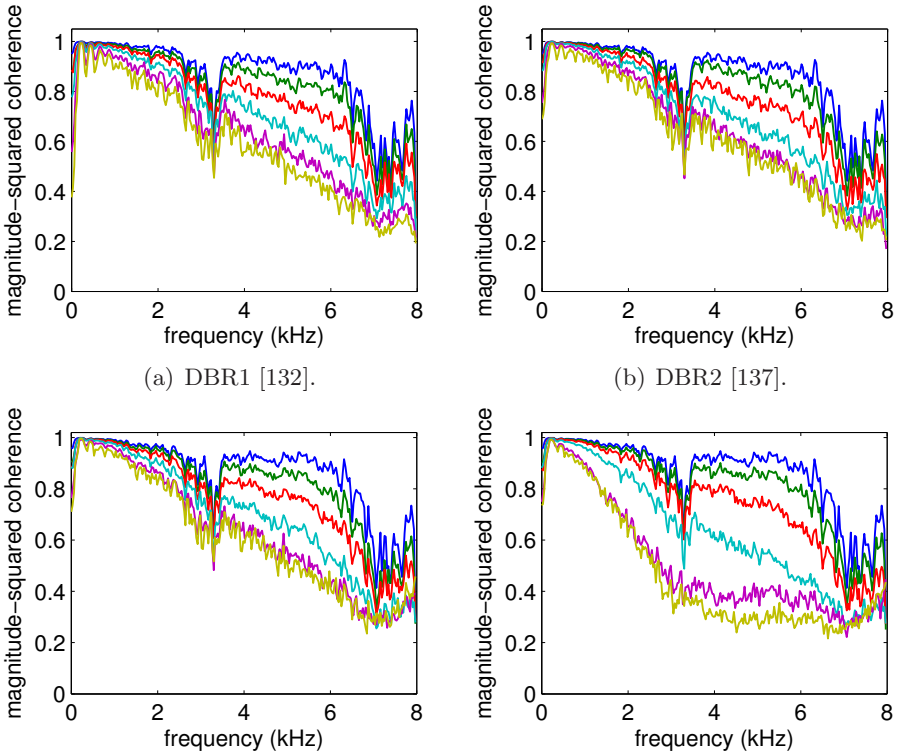
(d) DBR4.

Figure 26: Signal delay is linearly varied after resampling frame-wise (a) alternatingly across channels and (b,c,d) simultaneously across channels where every other block is resampled after time reversal and reversed back afterward ($N = 2048$, $R = 1.0004$, $f_s = 16$ kHz).

as possible across both time and channel. Also, increasing R for fixed N leads to a large reduction of the coherence from mid to high frequencies much like in Figure 27(d).

Figure 28 includes the result from nonlinear processing (NLP) with the nonlinearity parameter set at $\alpha = 0.5$ [34] for comparison with DBR. It show that NLP tends to reduce the coherence more than DBR at low frequencies where most of the speech energy resides. Depending on the frame size, DBR is able to reduce the coherence more than NLP at mid to high frequencies while altering the original signal less than NLP at low frequencies.

Informal listening tests have revealed that no audible distortion is noticed during loudspeaker playback after DBR as long as the loudspeakers are spaced sufficiently apart to eliminate the minor perception of "flutter," or audio image fluctuation, at high frequencies, most likely due to the coupling across frequency of time-varying phase difference between channels. The resampling rate R must be adjusted to be closer to unity in order to reduce the flutter effect for smaller frame size.



(a) DBR1 [132].

(b) DBR2 [137].

(c) DBR3.

(d) DBR4.

Figure 27: Inter-channel coherence (averaged over first 5 seconds of speech). From top to bottom: $N = \infty, 4096, 2048, 1024, 512, 256$.

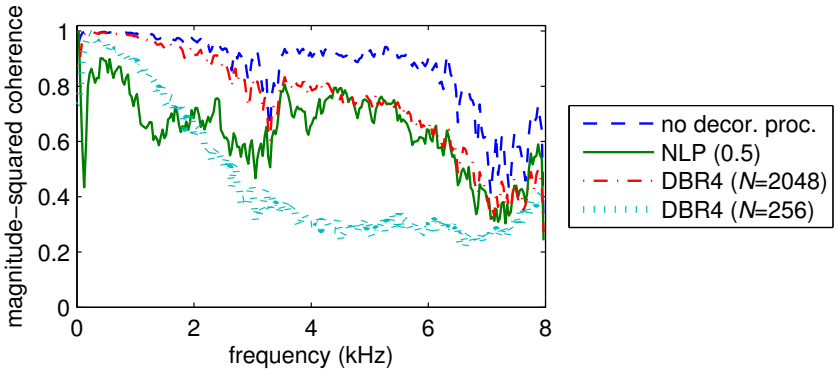


Figure 28: Inter-channel coherence comparison after decorrelation.

4.1.4 Decorrelation via Frequency-Domain Resampling (FDR)

As described above, DBR introduces time-varying signal delay frame-wise across channels with negligible audible distortion [132]. Resampling by interpolation (RBI) was utilized in [100, 132] since the conventional upsampling followed by downsampling may be computationally infeasible for real-time applications when the desired sampling rate change is very small (on the order of 0.01%). We may still do better computationally if we can take advantage of the efficiency of the FFT and resample, or interpolate, across frequency rather than across time with appropriate expansion or contraction dictated by the resampling ratio R.

Given the real-valued R in (89), the identity relationship

$$x(tR^{-1}) \longleftrightarrow X(e^{j2\pi fR}) \tag{92}$$

holds for continuous time signal $x(t)$ and the corresponding Fourier transform $X(e^{j2\pi f})$, *i.e.*, the time-frequency scaling is inversely related. Let $X(e^{j\omega})$ be the discrete-time Fourier transform (DTFT) and $X_N(k)$ be the N-point discrete Fourier transform (DFT) of $x(n)$, respectively, with proper bandlimiting during the sampling of $x(t)$ to avoid the frequency aliasing to obtain $x(n)$. Extending an N-point sequence from $x(n)$ by inserting zeros at the end of the sequence to turn it into an L-point sequence gives the relationship

$$\begin{aligned} y(n) &= \begin{cases} x(n), & n = 0, 1, \ldots, N-1, \\ 0, & n = N, N+1, \ldots, L, \end{cases} \\ &\longleftrightarrow \; Y_L(k) = U(e^{j\omega}) * X(e^{j\omega})\big|_{\omega = 2\pi k/L}, \end{aligned} \tag{93}$$

where "$*$" is the convolution operator and

$$\begin{aligned} u(n) &= \begin{cases} 1, & n = 0, 1, \ldots, N-1, \\ 0, & n = N, N+1, \ldots, L, \end{cases} \\ &\longleftrightarrow \; U_L(k) = U(e^{j\omega})\big|_{\omega = 2\pi k/L} = e^{-j\pi k(N-1)/L} \frac{\sin\big(\pi kN/L\big)}{\sin\big(\pi k/L\big)}. \end{aligned} \tag{94}$$

(93) specifies how often $Y_L(k)$, which results from the convolution (or smearing) of $X(e^{j\omega})$ with the sinc function over frequency due to the windowing of $x(n)$ by $u(n)$ over time, is sampled in the frequency domain. The phase shift $e^{-j\pi k(N-1)/L}$ in (94) is due to the DFT

of $u(n)$ not centered at the origin. On the other hand, expanding the N-point sequence of $x(n)$ by padding $M-1$ zeros between the samples gives an $L=MN$-point sequence with the relationships

$$y(n) = \begin{cases} x(n/M), & n = 0, M, 2M, \ldots, L-1, \\ 0, & \text{otherwise}, \end{cases}$$

$$\longleftrightarrow \; Y_L(k) = Y_N(k \bmod N) = X_N(k), \tag{95}$$

$$U_L^*(n) * y(n)$$

$$\longleftrightarrow \; X_L(k) = \begin{cases} X_N(k), & k = 0, 1, \ldots, N-1, \\ 0, & k = N, \ldots, L-1, \end{cases} \tag{96}$$

where "*" denotes the complex conjugation. (96) describes the M-times upsampling procedure with zero-padding followed by convolution with the sinc function for interpolation over time. The term $e^{j\pi n(N-1)/L}$ in $U_L^*(n)$ of (96) arises from the shifting of centered windowing, or low-pass filtering, in the frequency domain.

The above analysis implies that the DFT or the inverse-DFT of sampled signals can be interpolated automatically by zero-extending prior to the transformation. Specifically, increasing L by varying M with fixed N in (93) and (96) leads to resampled values over frequency and time, respectively. One must note, however, that the zero-extension technique does not increase the time or the frequency resolution, which is proportional to the number of non-zero samples N and not L, as no new information is added in the process.

Therefore, we propose the following procedure for resampling an N-point sequence from $x(n)$ by a factor of $0 < R < 2$:

1. Zero-extend the sequence by a factor of $M = 2^P$, $P \geq 1$.

2. Perform $L = MN$-point DFT on the extended sequence.

3. Linearly interpolate between k^{th} and $(k+1)^{th}$ samples

$$X'_L(k') = (1-\alpha')X_L(k) + \alpha' X_L(k+1) \tag{97}$$

with the constraints $k \leq Rk' \leq k+1$ and $\alpha' = Rk' - k$ for each $(k')^{th}$ new sample to form $2N$ equally spaced samples.

4. Perform $2N$-point inverse-DFT on the new samples.

5. Discard the samples at the end of the new sequence $x'(n)$ to retain the first RN resampled values, multiplied by the factor R.

Using the zero-extension factor $M \geq 2$ and taking the $2N$-point inverse-DFT avoids the time aliasing after resampling with $R > 1$. We assume M and N to be a power of 2 in general for efficient implementation of FFT. The computation load can be reduced further by taking advantage of conjugate-symmetric $X_L(k)$ for real-valued $x(n)$ and by storing the interpolated values over frequency in memory provided that they can be used later by the frequency-domain AEC.

Figure 29 illustrates the amount of distortion from resampling by the proposed FDR procedure when compared to the conventional time-domain resampling (TDR). It suggests that $4 \leq P \leq 6$ should be sufficient for most applications. FDR should offer computational saving over TDR when R is near unity and N is sufficiently small. For example, $2(128)+1$ interpolation filter coefficients for relatively high resampling accuracy were taken during TDR to obtain Figure 29, which translates to the computational overhead by roughly a factor of $257N/(MN\log_2(MN)) \approx 1.07$ when compared to FDR with $P = 4$ and $N = 2048$.

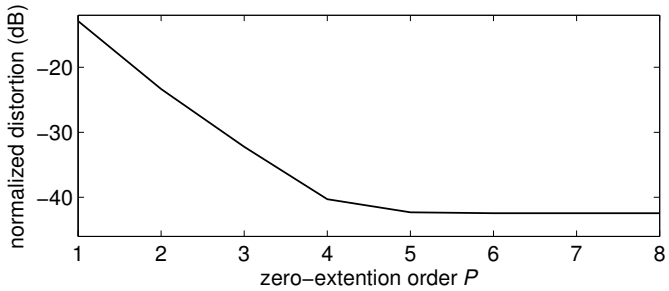


Figure 29: Distortion on a speech signal by FDR (normalized with respect to TDR) for $R = 1.0004$, $N = 2048$, and $f_s = 16$ kHz.

We note that one can choose to interpolate over the magnitude and the phase separately instead of over the real and the imaginary parts as implied by (97). It then necessitates the unwrapping of the phase, which may not be an accurate process for some signals. One can

also zero-extend at the beginning or evenly at both ends of a sequence rather than at the end. We have observed that applying these alternate approaches to a speech signal does not produce any perceptible difference. In any case, the overall distortion per resampled frame should be minimized as long as large enough P is taken to ensure accurate short-time interpolation. We also have observed that compared to FDR, TDR gives virtually the same coherence reduction as FDR except near 8 kHz where a notable roll-off in the coherence occurs due to the low-pass, anti-alias filtering. Thus FDR should provide even less frequency distortion than TDR.

4.1.5 Decorrelation via Sub-band Resampling (SBR)

For perceptual quality and the actual cancellation performance reasons [132, 137], we may want to modify the signal only in certain sub-bands. For example, interaural time differences plays an important role for sound localization at low frequencies [140]. A modification of the low-band content disturbs the phase information of the signal and ultimately alters the interaural time differences.

To that end, we point out that for achieving the same overall reduction in the coherence, or equivalently the cross-correlation, the resampling ratio R may be adjusted separately over each sub-band in the frequency domain as if resampling the entire signal frame with a fixed R. This can be done to make sure that the spatial image distortion will be minimized by the resampling process. In addition, a sudden change in R between sub-bands, *e.g.*, $R = 1$ in the low band and $R = R_0 > 1$ in the high band, may introduce the unwanted frequency-domain distortion. We have experimentally verified that the distortion created by such a discontinuity in R has the characteristics of a musical noise. Therefore, we propose to vary the resampling ratio per frequency bin as smoothly across the bins as possible, which simply involves making R a continuous function of frequency, *i.e.*, $R(k)$, and applying the desired $R(k)$ curve to the FDR procedure described previously.

Unlike the traditional decorrelation procedures [34, 125, 51] where the coherence is usually fixed throughout the frequencies for a given parameter, SBR is highly flexible and can be fine-tuned for better perceptual quality, *e.g.*, less "resampling" at lower frequencies and

vice versa at higher frequencies. We have shown that given the same degree of decorrelation, SBR outperforms the conventional methods in terms of the audio quality [141].

4.2 Extension of Robust AEC System

4.2.1 Robust MCAEC

Just as in [131], FBLMS can be combined with the REE procedure for MCAEC by using a "compressive" nonlinearity for the ERN, which provides robustness against an impulsive noise and permits continuous adaptation during double talk, and with the regularized normalization factor [52]

$$\frac{S_{x_j}(k,l)}{S_{x_j}^2(k,l) + \gamma S_{v_i}^2(k,l)}, \tag{98}$$

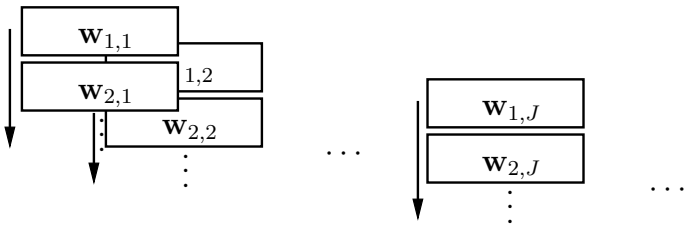
which ensures a wide range of noise robustness (*e.g.*, when the echo path is weakly excited) and is an integral part of the REE procedure, where the reference and the noise power spectrums S_{x_j} and S_{v_i} for k^{th} frequency bin at l^{th} block index are determined per j^{th} and i^{th} channels, respectively. The same simplified statistics estimation strategy in [131] can be utilized, where S_{v_i} in (98) is estimated directly by the residual echo power spectrum S_{e_i} and the over-suppression factor $\eta \geq 1$ is employed this time with REE such that the signal-to-noise ratio (SNR) $= \eta$ for improved stability ($\eta = 1$ in [131]).

The following modifications may be included for extra improvement in the overall MCAEC performance. First, double-talk detection (DTD) [39] is used to decrease the step-size by half during double talk to maintain as much stability as possible. Second, exponential weighting (EW) [72] is applied to the time-domain filter coefficients during the FBLMS's gradient constraint procedure for increased convergence rate and also adaptation stability. Finally, a truly batch-wise adaptation can be carried out for a batch of B samples, where the filtering and the adaptation steps are performed per block size L (same as the filter size) and repeated across blocks of $L < B$, with or without overlap, in the same batch for *iter* iterations ($B = L$ in [131]).

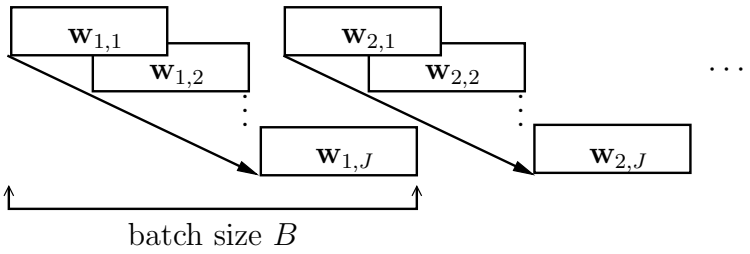
4.2.2 Variations in BIA

There are mainly two options for BIA implementation. BIA1 in Figure 30(a) adapts over each filter block, assumed here to be of size L, for several iterations before moving on to

the next block in time. BIA2 in Figure 30(b) adapts across J blocks partitioned over a batch of data of size B, and the process is repeated for several iterations. The blocks may be overlapped over time with the integer overlap factor $of > 1$ ($of = 2$ in Figure 30) for further increase in the convergence speed, which then requires proper output smoothing [83] [1]. The two cases are each suited for real-time and off-line implementations, respectively. If the adaptation is permitted for $B > L$, BIA2 should provide higher tERLE than BIA1 for the same amount of calculation since it allows for wider input signal variation in time, which is beneficial for a correlated and non-stationary signal such as speech.



(a) BIA1: Adaptive iteration per block.



(b) BIA2: Adaptive iteration across blocks.

Figure 30: Block-iterative adaptation (BIA) of the filter coefficient vector $\mathbf{w}_{i,j}$, where i is the iteration index and j is the block index.

4.2.3 Multi-delay Filter (MDF)

The MDF partitions $\mathbf{x}$ and $\mathbf{w}$ into K sub-blocks of size D each for reduced computation in the discrete Fourier transform (DFT) domain:

$$\hat{d}(n) = \mathbf{w}^T(n)\mathbf{x}(n) = \sum_{k=0}^{K-1} \mathbf{w}_k^T(n)\mathbf{x}_k(n), \tag{99}$$

[1] Smoothing of the filter output $\hat{d}$ for $of > 1$ as suggested in [83] actually interferes with BIA. It is sufficient to simply smooth the residual echo e to eliminate the audible discontinuity due to the block-adaptive processing.

where $\mathbf{w} = [\mathbf{w}_1^T, \ldots, \mathbf{w}_K^T]^T$, $\mathbf{x} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_K^T]^T$, and $L = KD$ is the filter length. On the other hand for MCAEC with P channels, the acoustic echo estimate at one of the microphones is given by

$$\hat{d}(n) = \mathbf{w}^T(n)\mathbf{x}(n) = \sum_{i=0}^{P-1} \mathbf{w}_i^T(n)\mathbf{x}_i(n), \tag{100}$$

were $\mathbf{w} = [\mathbf{w}_1^T, \ldots, \mathbf{w}_P^T]^T$ and $\mathbf{x} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_P^T]^T$ are the concatenated filter coefficient and reference signal vectors, respectively. We can observe from (99) and (100) that the inter-block and the inter-channel correlation problems are due to high correlation between the sub-blocks of the reference vector $\mathbf{x}$. Thus we should expect the same systematic benefit from applying REE and BIA to MDF as when they are applied to MCAEC.

Compared to the full-block version, FBLMS, the advantages and the disadvantages of the MDF algorithm are as follows:

Advantage (*computation-wise*):

- Shorter DFT size and blocking delay.
- Reuse of prior DFT($\mathbf{x}_k$) stored in memory.
- Error enhancement over shorter error block $\mathbf{e}_k$.
- Less BIA iterations necessary for blocks towards the RIR tail with less energy.

Disadvantage (*performance-wise*):

- Shorter $\mathbf{x}_k$ means increased ill-conditioning of $\mathbf{R}_{x_k x_k}$ for a speech signal, thus better regularization is required.
- Shorter $\mathbf{e}_k$ is used to update all filter coefficients, thus better error enhancement is required.

The generalized MDF (GMDF), *i.e.*, MDF with $of > 1$, allows for increase in the convergence speed [83]. In such a case, BIA2 in Figure 30(b) may be executed for each sub-block (*i.e.*, $J = of$ and $B = \frac{of+1}{of}D$) for a gain in the tERLE.

4.3 Experimental Evaluation

TIMIT speech corpus recorded at 16 kHz sampling rate was used to simulate the near and far-end talkers with at most two simultaneous talkers at both ends with an overlap of at most 2 seconds (see Figure 31). Same number of microphones, loudspeakers, and talkers was used at each end. Talkers were randomly selected such that the talkers from both ends took turns to speak exactly one utterance per sequence, and such a sequence was repeated for at least 20 seconds for each simulation. The signal energy during voice activity after decorrelation was normalized to match that of the original signal to maintain the same energy after all decorrelation procedures and ensure consistent echo return loss (ERL) control.

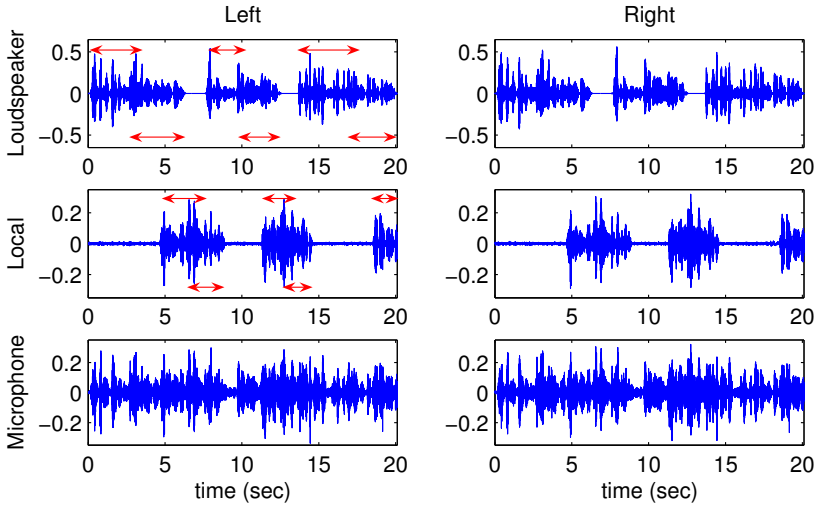


Figure 31: Near-end loudspeaker, local, and microphone signals for SAEC. Double-arrows indicate individual speech activity.

4.3.1 Robust MCAEC with and without DBR

The following decorrelation procedures were tested with simulated SAEC to initially evaluate the DBR technique.

- Nonlinear processor (NLP) [34].
- Additive white Gaussian noise (AWGN).
- One-sample delay (OSD) [125].

- Resampling by upsampling and downsampling (RUD) via DBR1.
- Resampling by interpolation (RBI) [100] via DBR1.

In order to eliminate the audible "pops" generated between the processed frames due to discontinuity for the last three methods, the following simplified smoothing schemes were used, where $x_j(m,n)$ and $\hat{x}_j(m,n)$ are the original and the resampled values, respectively, of n^{th} sample in m^{th} (current) frame of size N_f from j^{th} loudspeaker channel.

- OSD: The average of $x_j(m-1,N_f)$ and $x_j(m,1)$ is inserted between the two samples to create one-sample delay. In order to avoid the accumulation of delay, $x_j(m,N_f)$ is overlapped and averaged with $x_j(m+1,1)$.
- RUD: $R>1$ is chosen such that one extra sample is produced by resampling. Afterward, $\hat{x}_j(m,1)$ is averaged with $x_j(m-1,N_f)$, and the extra sample $\hat{x}_j(m,N_f+1)$ is overlapped and averaged with $x_j(m+1,1)$.
- RBI: After resampling-by-interpolation that produces the same number of samples as before, $\hat{x}_j(m,1)$ is averaged with $x_j(m-1,N_f)$, and $\hat{x}_j(m,N_f)$ is averaged with $x_j(m+1,1)$.

Only one extra sample delay (look-ahead) is incurred by the above strategies for real-time playback. More advanced framing and smoothing are possible, *e.g.*, [125], albeit with longer delay.

A pair of omni-directional microphones 2 cm apart were placed 50 cm away from the middle of a pair of loudspeakers 50 cm apart ($P=Q=2$). The configuration was used to record three sets of RIRs, two for the near and far-end talkers and one for the near-end loudspeakers, with average reverberation time of $T_{60}=250$ ms. The RIRs were truncated to 128 ms ($L=2048$ at $f_s=16$ kHz) before convolution, and the near-end RIRs were scaled to produce the ERL of 10 dB. 40 dB SNR AWGN was applied to the far-end microphone signals, and an air-conditioner noise and local speech signals (double talk) with echo-to-noise ratios (ENRs) of 20 dB and 0 dB, respectively, were mixed with the acoustic echo to

comprise the near-end microphone signals. $\mu = 0.15$, $\beta = 0.99$, $\gamma = 1$, and $\eta = 5$ were used for FBLMS and REE [132].

Figure 32 indicates that BIA accelerates the convergence rate significantly especially at the beginning of adaptation (in contrast to the common approach of simply using a fast-converging adaptive algorithm to combat the effect of the non-uniqueness problem). Using DTD to decrease the step-size during double talk assists in increasing the tERLE calculated without the local noise. The effectiveness of EW during both single and double talk is also observed. Continuous, noise-robust adaptation afforded by REE is crucial for MCAEC since the far-end room response change may occur during double talk, *e.g.*, far-end speech activity switches at $t \approx 7.5$ seconds in Figure 31.

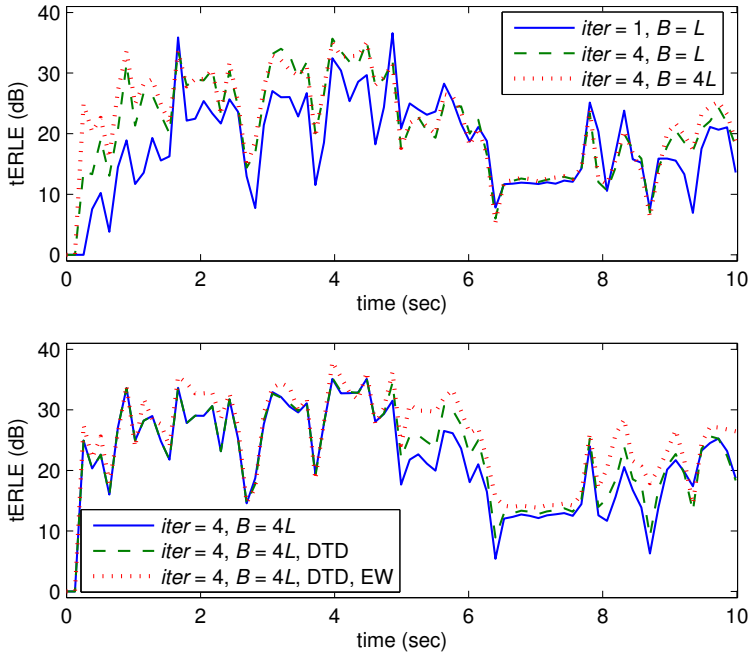


Figure 32: True ERLE (averaged over left and right channels) for combinations of *iter*, B, DTD, and EW without decorrelation.

However, Figure 33 reveals that only a minor improvement in the overall misalignment is possible without a decorrelation procedure. The effect of the far-end speech activity transition is clearly visible at $t \approx 2.5$ seconds in Figure 34 without decorrelation even when all other techniques are employed. Some improvement is displayed after decorrelation in Figures 35, 36, and 37 for NLP (the nonlinearity parameter was set at $\alpha = 0.5$ [34]), AWGN

(30 dB SNR), and OSD ($N_f = L$), respectively, but with limitations, *e.g.*, degradation of the near steady-state performance by NLP is quite apparent.

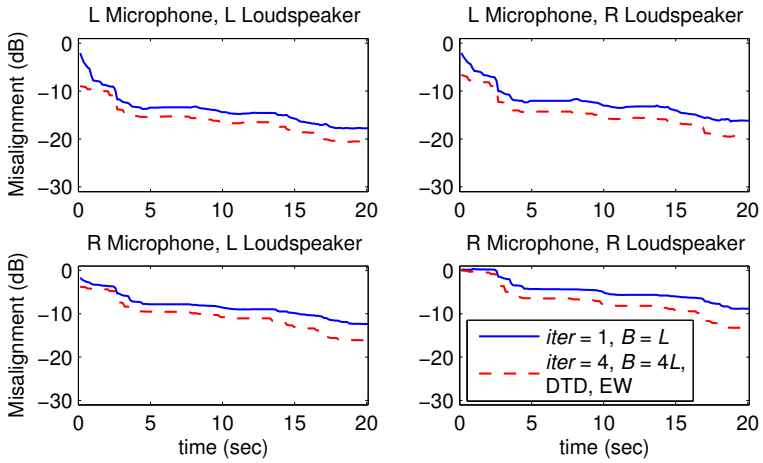


Figure 33: Improvement in misalignment without decorrelation.

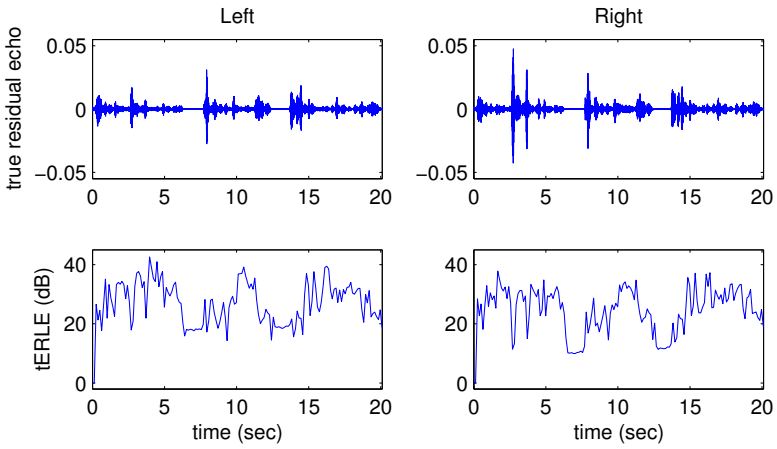


Figure 34: True residual echo and tERLE without decorrelation.

On the other hand, Figures 38 and 39 confirm the effectiveness of the resampling technique. RBI provides virtually the same results as RUD (`MATLAB`'s resample function was used for RUD, the reuse-block size and sinc function length of 64 was used for RBI [100], and $N_f = L$ and $R = 1.0004$ were used for both). Informal listening tests indicated no loss in perceptual quality after decorrelation by RUD or RBI, whereas remaining distortion in the residual echo was obvious for NLP and AWGN.

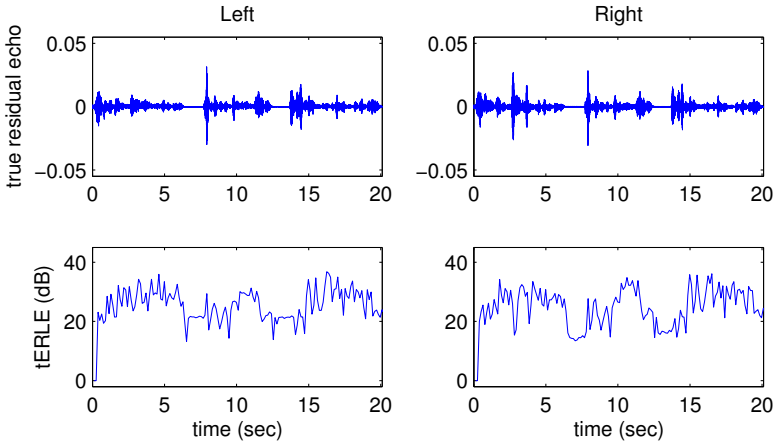


Figure 35: True residual echo and tERLE with NLP.

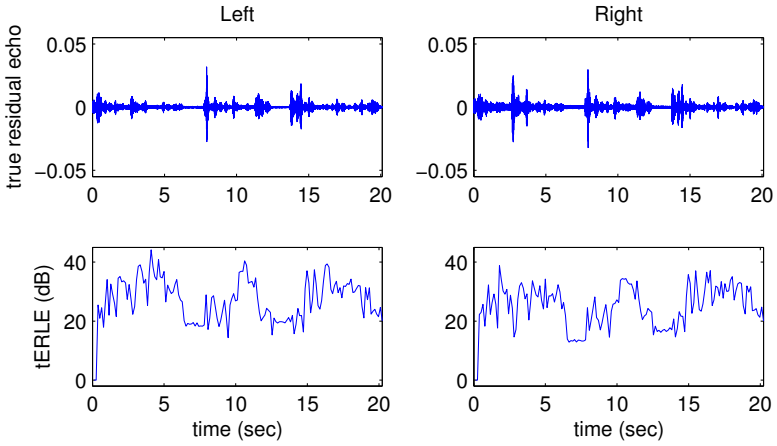


Figure 36: True residual echo and tERLE with AWGN.

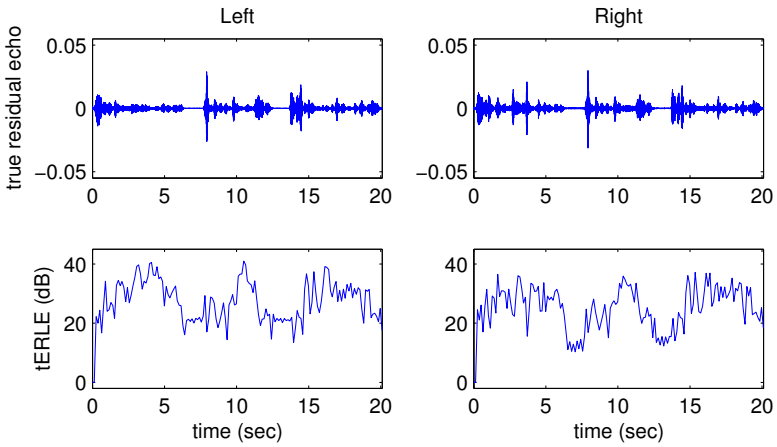


Figure 37: True residual echo and tERLE with OSD.

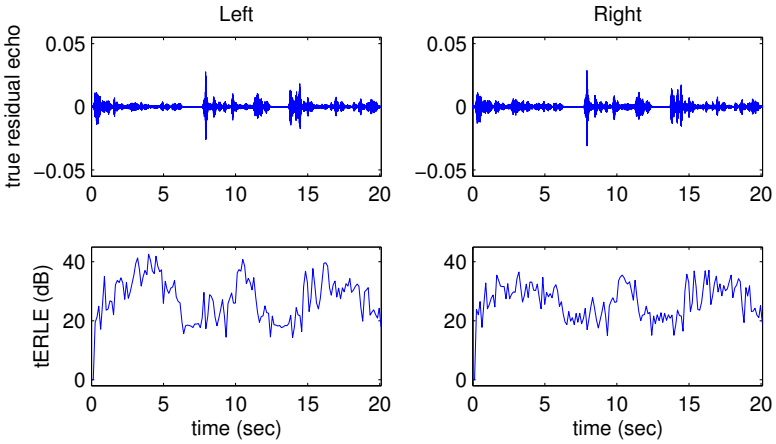


Figure 38: True residual echo and tERLE with RUD or RBI.

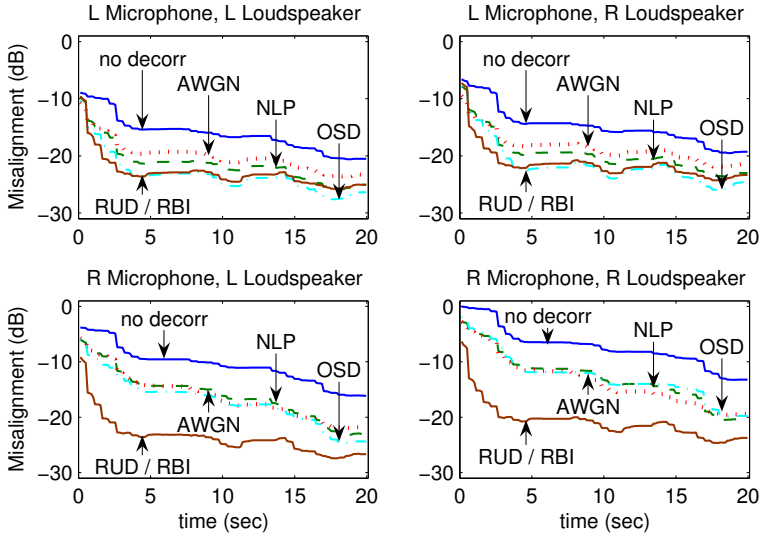


Figure 39: Improvement in misalignment after decorrelation.

4.3.2 Robust MCAEC with FDR

The following cases were tested for MCAEC to evaluate the FDR technique.

- NLP, modified to work for more than 2 channels as

$$\tilde{x}_i(n) = x_i(n) + \frac{\alpha}{2}\left(x_i(n) + (-1)^{\bmod(i-1,2)}|x_i(n)|\right), \tag{101}$$

  here $x_i(n)$ is the reference signal from i^{th} channel, $i \geq 1$, and $\bmod(\cdot,\cdot)$ is the modulus function.

- AWGN.

- FDR via DBR2, DBR3, or DBR4.

For the rest of the simulations in this chapter, a set of RIRs with the reverberation time of $T_{60} \approx 200$ ms was measured using an 8-microphone (omni-directional) circular array with 0.02 m spacing placed in the center of an 8-loudspeaker circular array with 1 m spacing. The arrays were ordered counter-clockwise, and the RIRs between four microphones with indicies $i \in \{1,2,3,4\}$ and four loudspeakers with indicies $j \in \{1,2,3,4\}$ were used for the near-end echo paths, those between $i \in \{1,2,3,4\}$ and $j \in \{5,6,7,8\}$ for the near-end speech mixing, and those between $i \in \{5,6,7,8\}$ and $j \in \{5,6,7,8\}$ for the far-end speech mixing. Different sets of loudspeakers were used for the sets $\{1,2,3,4\}$ and $\{5,6,7,8\}$ to vary the near and the far-end echo paths as much as possible. The RIRs were truncated to 128 ms before convolution to set it equal to the filter length ($L = 2048$ at $f_s = 16$ kHz), and the near-end RIRs were scaled to produce the ERL of around 10 dB.

4.3.2.1 Wideband tERLE and Misalignment Evaluation

40 dB and 100 dB SNR AWGNs were applied to the near and far-end microphones, respectively, and a background noise from an air-conditioner and local speech signals with the ENRs of 20 dB and 0 dB, respectively, were mixed with the acoustic echo to comprise the near-end microphone signals. $\mu = 0.02$, $\beta = 0.998$, $\gamma = 1$, $\eta = 1$, $B = L$, and $iter = 4$ along with DTD (to decrease the step-size by half during double talk) and EW were used for the REE-based FBLMS algorithm [137]. Smaller step-size μ was used this time than in

[132] to achieve better steady-state performance at the cost of slower convergence speed. For decorrelation, NLP was used with $\alpha = 0.5$, AWGN was generated at 30 dB SNR, and FDR was used with $P = 6$, $R = 1.0004$, and $N = 2048$.

Tables 3, 4, 5 provide the tERLE, segmental signal-to-residual echo ratio (SSRR, calculated in the same fashion as the segmental SNR between v and b), and LSD (measured between v and e), respectively. The measurements were averaged over all channels for 20 independent simulations, where the last 19 out of 20 seconds of data were used to ensure sufficient initial filter convergence. SSRR and LSD quantify the amount of time-domain and frequency-domain distortions, respectively, remaining after the echo cancellation process. tERLE was measured during voice activity only. The tables show that FDR is able to maintain consistently better performance than NLP or AWGN as the number of channels is increased, although the apparent differences are relatively small due to the averaging process. In addition, as already referred to in [132], all decorrelation procedures tend to result in improved tracking of the echo paths but at the cost of decreased initial and steady-state tERLE. The trade-off is more pronounced when a larger step-size is used. Therefore, retaining the original reference signal's characteristics during decorrelation can be beneficial for the best overall cancellation performance.

Figure 40 provides the misalignment plot, averaged over all echo paths, that indicates the advantage of FDR over others for quicker tracking performance. Block-iterative adaptation, which may be generalized as batch-wise adaptation or data reuse, is essential to frequency-domain MCAEC since it naturally allows the recovery of tERLE lost due to the non-uniqueness problem [132]. Data reuse by the normalized LMS (NLMS) algorithm is shown to be another form of the affine projection algorithm (APA)[8], and BIA of the FBLMS algorithm evidently exhibits an APA-like convergence behavior [137]. For MCAEC, a decorrelation procedure should not only reduce the coherence across channels but also emphasize the natural variation of the correlation across time such that a projection-type adaptive algorithm may "zig-zag" its way faster towards the optimal solution than otherwise [110]. Stabilized batch adaptation of the FBLMS algorithm through the REE procedure should be compatible with the time-varying decorrelation procedure via resampling since

Table 3: tERLE comparison (dB, higher is better).

# channels	*none*	NLP	AWGN	FDR
2×2	*20.5*	19.7	20.0	**20.5**
3×3	*18.9*	18.4	18.4	**19.0**
4×4	*18.1*	17.5	17.5	**18.0**

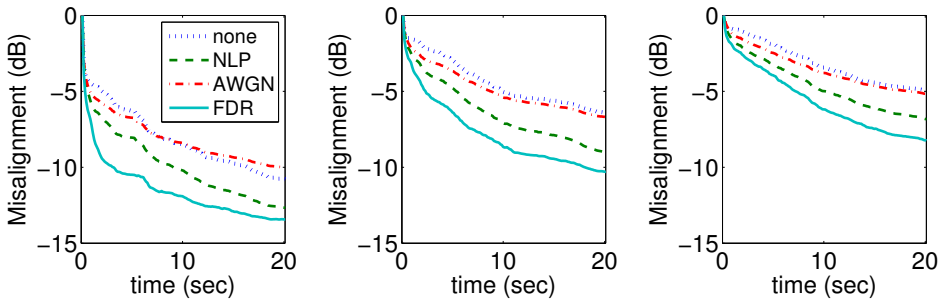
Table 4: SSRR comparison (dB, higher is better).

# channels	*none*	NLP	AWGN	FDR
2×2	*20.3*	19.7	19.6	**20.3**
3×3	*19.8*	19.3	18.8	**19.7**
4×4	*19.7*	19.3	18.6	**19.6**

Table 5: LSD comparison (lower is better).

# channels	*none*	NLP	AWGN	FDR
2×2	*0.657*	0.727	0.700	**0.648**
3×3	*0.723*	0.772	0.779	**0.706**
4×4	*0.785*	0.848	0.854	**0.783**

less decorrelation is then required for the low-frequency signal components that carry most of the speech energy and information. This in turn minimizes the interference of the actual linear cancellation process and ultimately aids in the maximization of the cancellation performance in a complex real-world environment, where low misalignment is only a sufficient and not a necessary condition for high tERLE [137] especially for MCAEC with multiple echo paths in a very noisy condition.



(a) 2×2 case. (b) 3×3 case. (c) 4×4 case.

Figure 40: Misalignment averaged over 20 runs and all echo paths.

4.3.2.2 Sub-band tERLE and Misalignment Decomposition

The tERLE and the misalignment were decomposed into three sub-band components (low, mid, and high) through the Fourier series expansion, which gives far better reconstruction accuracy than using the DFT filter banks formed from a prototype filter. For tERLE, the microphone and the residual echo signals were decomposed with 50% overlap of the analysis frames. For misalignment, the RIR and the filter coefficients were mirrored in time to extend their size by a factor of two prior to the decomposition.

This time the near-end RIRs were switched at 15 sec to enact a sudden disruption to the RIR. 100 dB and 40 dB SNR AWGNs were added to the near and far-end microphones, respectively, and air-conditioner noise and speech with the ENRs of 20 dB and 0 dB, respectively, were added to the acoustic echo. The the REE-based FBLMS algorithm used the parameters $\mu = 0.12$, $\beta = 0.998$, $\gamma = 10$, $\eta = 5$, $iter = 4$, $B = L$ (*i.e.*, BIA1 in Figure 30(a)), and $of = 4$ ($of = 1$ was used in [132, 137]) along with EW and scaling of the step-size μ by half during double talk [132]. For decorrelation, NLP was used with $\alpha = 0.5$ and FDR was used with $P = 6$ and with DBR4 of given frame size.

Figures 41, 42, and Table 6 show the SAEC results. The averaged tERLE was obtained over the echo duration only while the misalignment was averaged over the entire time. The far-end talker activity change takes place four times, occurring initially at around 2.5 sec. The main observations are as follows. First, both NLP and DBR tend to hurt the initial and the steady-state tERLE for improved echo-path tracking. Second, NLP does not fare as well as DBR in the mid and high bands after the far and the near-end RIR change. For the low band, NLP leads to lower steady-state misalignment but not higher tERLE than DBR1. Over all bands, DBR is able to give higher tERLE and lower misalignment on average than NLP. Third, a substantial gain in the tracking capability appears in the mid to high bands for DBR. Such an improvement is attributed largely to the DBR's ability to continuously instill both short and long-time decorrelation for the direct benefit of the LMS algorithm and not simply due to the coherence reduction. Finally, smaller N leads to much better tracking for DBR, where the two features are both crucial for real-time AEC.

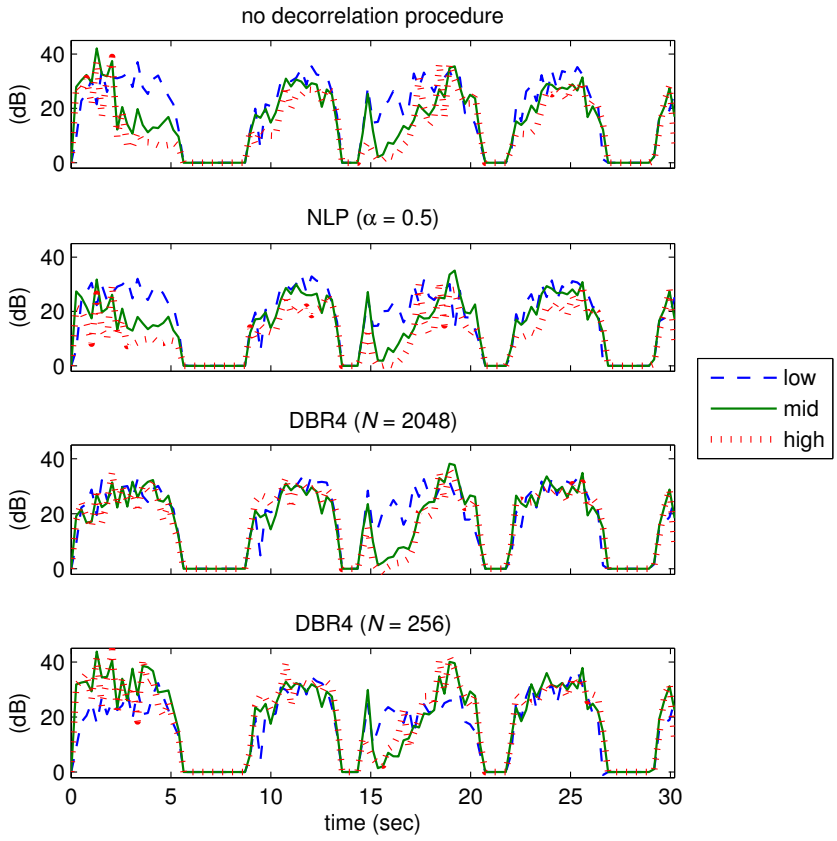


Figure 41: Sub-band tERLE decomposition (stereophonic, averaged over all channels).

Table 6: Average tERLE (dB, left column) and misalignment(dB, right column).

band	*no decor.*		NLP		DBR4$_{N=2048}$		DBR4$_{N=256}$	
low	*24.6*	*-8.1*	22.6	-9.8	**23.7**	-9.6	22.8	**-11.1**
mid	*20.4*	*-10.7*	19.3	-13.4	22.3	-21.7	**25.3**	**-23.9**
high	*17.2*	*-7.8*	15.2	-10.3	22.0	-20.8	**25.2**	**-22.2**
all	*24.3*	*-8.3*	22.9	-10.5	**24.3**	-15.8	23.9	**-17.1**

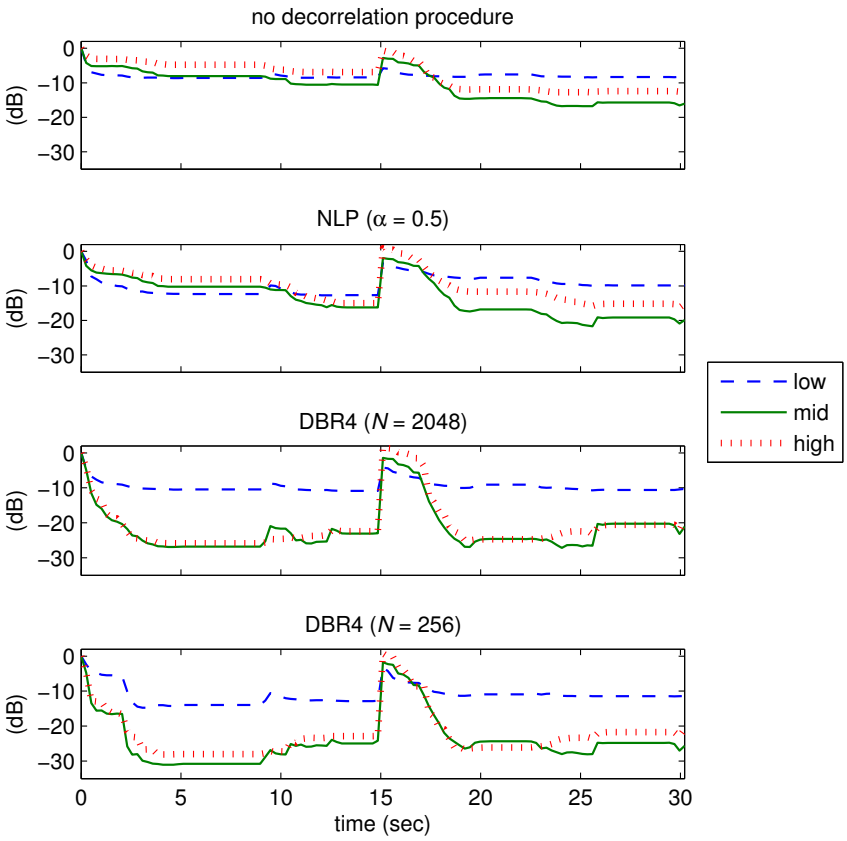


Figure 42: Sub-band misalignment decomposition (stereophonic, averaged over all echo paths).

4.3.3 Robust MCAEC with SBR

The following methods were tested to compare the R-AEC performance with the proposed decorrelation by SBR against other commonly used decorrelation procedures.

- AWGN at 15 dB segmental signal-to-noise ratio (SSNR).
- NLP with $\alpha = 0.5$ [34].
- Phase modulation (PMod) proposed by [51].
- FDR via DBR3 with fixed resampling ratio $R = 1.0028$.
- SBR via DBR3 with $N = 512$ and variable resampling ratios R_1 through R_5 as shown in Figures 43 and 44.

4.3.3.1 Quality Evaluation

For the evaluation of speech quality after decorrelation, a stereo reference signal of 30 seconds was used. Silences were removed prior to calculating the coherence. As SBR allows us to fine-tune the coherence at each frequency bin, R_1 is used to achieve the same coherence given by AWGN, R_2 to achieve that by NLP, and R_3 to achieve that by PMod to form the same basis for measuring the processed speech quality and comparing against other decorrelation procedures. Figure 43 also shows how well the coherence can be controlled by SBR. Thus by properly choosing $\Delta R = R - 1$, the average degree of decorrelation, measured in terms of the coherence, by SBR can be matched to that of AWGN, NLP, and PMod. Also to demonstrate the ability of SBR to control the AEC performance per sub-band, the coherence is matched with regular FDR only in the mid to high bands while leaving the low band unmodified. The two other SBR coherence-matching schemes with the variable resampling ratios R_4 and R_5 used for this purpose are shown in Figure 44.

For objective quality evaluation, SSNR, LSD, and perceptual evaluation of speech quality (PESQ) score were used. The SSNR measures the deviation of the processed signal from the original signal in the time domain while the LSD measures the distortion in the frequency domain. Both narrowband and wideband modes were used for the PESQ score

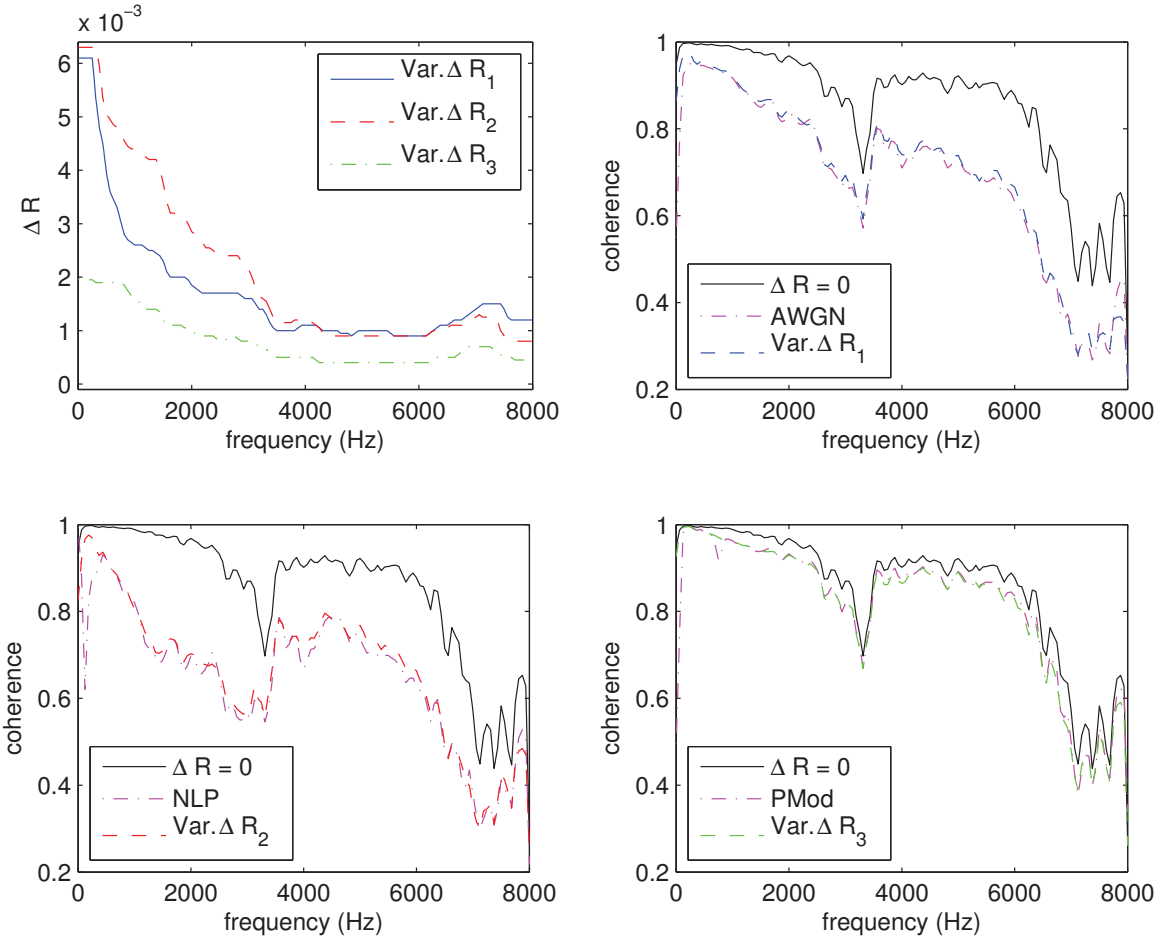


Figure 43: Variable resampling ratios R_1, R_2, and R_3 and their corresponding coherence plots, which match the coherence from SBR to that of other decorrelation methods.

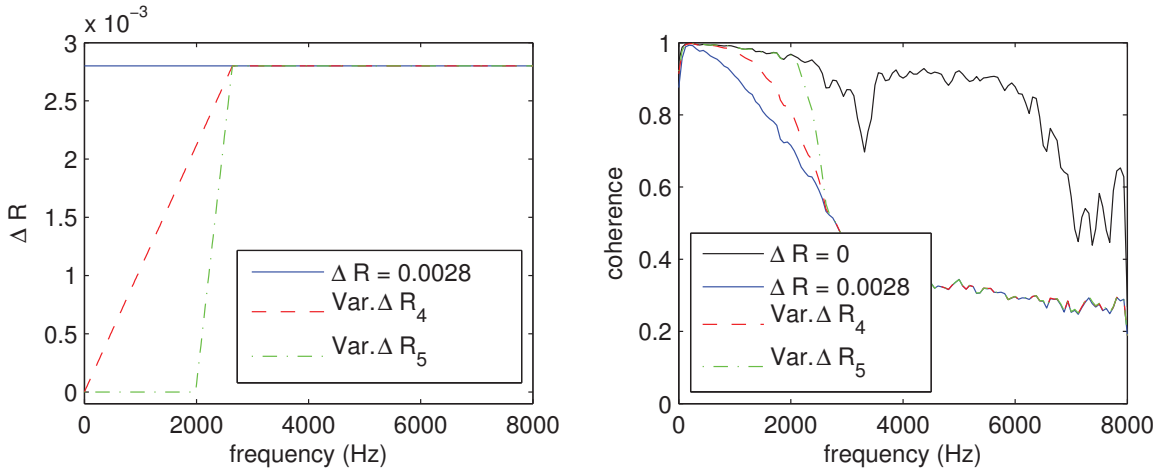


Figure 44: FDR with fixed $\Delta R = 0.0028$ and the corresponding variable resampling ratios R_4 and R_5 that match the coherence of FDR in the high frequency band.

(PESQ$^{\text{NB}}$ and PESQ$^{\text{WB}}$), which is an objective measurement that predicts the results of mean opinion score (MOS) in subjective listening tests. PESQ$^{\text{NB-LR}}$ and PESQ$^{\text{WB-LR}}$ correspond to the evaluations obtained after averaging the measures taken individually from the left and the right channels.

Table 7 summarizes the quality measures. SBR generally outperforms the other conventional methods in terms of the sound quality as reflected by the PESQ score [141]. We note that even though the SSNR and the LSD of AWGN are better than R_1, the distortion introduced by AWGN is quite audible as indicated by the PESQ score. The distortion introduced by SBR, on the other hand, is almost negligible when ΔR is very small. We also note that the resampling ratio for FDR and R_4 and R_5 for SBR in the igh band are set to large values to demonstrate the effect of highly decorrelated signal after DBR on the tERLE and the misalignment performances. As a result, the PESQ$^{\text{WB}}$ score suffers due to large distortion in the high frequency region. Still, even though the PESQ scores are quite similar in those cases, better SSNR and LSD can be achieved by avoiding the resampling of the low band as reflected by SBR with R_5.

Table 7: Processed speech quality comparison.

method	AWGN	R_1	NLP	R_2	PMod	R_3	FDR	R_4	R_5
SSNR	**15.00**	8.68	3.24	**7.22**	1.48	**13.62**	6.62	8.39	**9.51**
LSD	**0.07**	0.51	2.45	**0.66**	0.37	**0.24**	**0.79**	0.66	0.59
PESQ$^{\text{NB-LR}}$	3.67	**4.50**	4.03	**4.49**	4.53	**4.53**	4.52	4.55	**4.55**
PESQ$^{\text{NB}}$	3.57	**4.51**	4.39	**4.51**	3.94	**4.55**	**4.25**	4.24	4.24
PESQ$^{\text{WB-LR}}$	3.56	**4.59**	3.76	**4.57**	4.62	**4.63**	4.61	4.63	**4.63**
PESQ$^{\text{WB}}$	3.52	**4.48**	3.96	**4.47**	2.56	**4.63**	3.73	3.74	**3.74**

4.3.3.2 Sub-band tERLE and Misalignment Decomposition

Just as in the case of FDR, the REE-based FBLMS algorithm was used with $\mu = 0.12$, $\beta = 0.998$, $\gamma = 10$, $\eta = 5$, $iter = 4$, $B = L$, and the overlap factor $of = 4$ ($of = 1$ was used in [132, 132]) along with EW and scaling of the step-size μ by half during double talk [132].

Tables 8 and 9 show the SAEC results. The main observations are as follows. First, although AWGN is able to provide the tERLE closer to when no decorrelation is used than SBR, it leads to much worse misalignment. Second, NLP tends to hurt the low-band tERLE

more than SBR when compared to no decorrelation. The performance gain by SBR against NLP is even larger in the mid and the high bands for the tERLE and especially for the misalignment. Finally, PMod is capable of providing lower misalignment over all bands than SBR when its coherence is matched by SBR, but it does not necessarily translate to higher tERLE, which is less in the low to mid bands for PMod than SBR. Poor misalignment by SBR in this case is expected since the coherence is not reduced much after the matching.

Table 8: Average tERLE (dB, higher is better).

band	*none*	AWGN	SBR R_1	NLP	SBR R_2	PMod	SBR R_3
low	*24.5*	**24.4**	23.3	22.6	**23.1**	21.5	**24.1**
mid	*20.1*	19.8	19.8	18.7	**19.7**	18.7	**19.4**
high	*17.1*	15.5	**17.5**	14.3	**17.3**	**17.3**	15.8
all	*23.8*	**23.7**	22.9	22.2	**22.7**	21.1	**23.3**

Table 9: Average misalignment (dB, lower is better).

band	*none*	AWGN	SBR R_1	NLP	SBR R_2	PMod	SBR R_3
low	*-8.1*	-8.0	**-9.5**	-9.9	**-10.1**	**-9.6**	-8.5
mid	*-10.9*	-11.1	**-16.8**	-13.6	**-16.8**	**-18.7**	-12.1
high	*-7.9*	-8.3	**-15.7**	-10.5	**-15.3**	**-15.5**	-9.8
all	*-8.4*	-8.7	**-13.7**	-10.7	**-13.8**	**-13.6**	-9.9

The results in Tables 10 and 11 (averaged only over echo duration for tERLE and over entire time for misalignment) indicate that a substantial gain in the tracking capability appears in the mid to high bands for FDR and SBR when compared to no decorrelation and other decorrelation procedures. The tERLE is also increased especially in the high band. Such an improvement is again attributed to the DBR's ability to continuously instill both short and long-time decorrelation for the benefit of the LMS algorithm. Furthermore, SBR is able to provide higher tERLE than FDR in the low band by selectively not modifying the signal components in that region, in which case the tERLE is recovered naturally through BIA of the REE technique. This results in higher overall tERLE for SBR than without decorrelation.

Table 10: Average tERLE (dB, higher is better).

band	*none*	FDR	SBR R_4	SBR R_5
low	*24.5*	23.6	24.2	**24.3**
mid	*20.1*	21.7	21.7	21.7
high	*17.1*	21.3	21.3	21.3
all	*23.8*	23.7	24.2	**24.3**

Table 11: Average misalignment (dB, lower is better).

band	*none*	FDR	SBR R_4	SBR R_5
low	*-8.1*	**-9.4**	-8.7	-8.1
mid	*-10.9*	-21.7	-21.7	-21.7
high	*-7.9*	-21.0	-21.0	-21.0
all	*-8.4*	**-15.8**	-15.2	-14.8

4.3.4 Robust MDF

The system parameters used for FBLMS were $\mu = 0.12$, $\beta = 0.998$, $\gamma = 10$, $\eta = 5$, $iter = 4$, $B = L$ (*i.e.*, BIA1 in Figure 30(a)), and $of = 4$ ($of = 1$ was used in [132, 137]) along with EW and scaling of the step-size μ by half during double talk [132]. For MDF with $K = 4$, the parameters adjusted from FBLMS were $\mu = 0.35$, $\gamma = 10K^2$, $\eta = 10$, and $of = 1$. For GMDF, BIA2 in Figure 30(b) was applied with $\mu = 0.2$, $\gamma = 20K^2$, $\eta = 10$, and $J = of = 4$.

Figures 45, 46, and Table 12 provide the results from single-channel AEC. The figures show that, as BIA with $iter = 4$ is used instead of the usual $iter = 1$, MDF is capable of closely matching the performance of FBLMS even in a noisy condition. Due to the increased overlap, GMDF is able to provide better result than MDF. Still, the MDF is limited by shorter DFT blocks than with FBLMS, which affects the high frequency components more than the lower ones. Further improvement is expected through refined regularization and error enhancement procedures, *e.g.*, [143]. Although not plotted here, the reduction of BIA in each sub-block according to $iter = 4, 3, 2, 1$ for $k = 1, 2, 3, 4$, respectively (*i.e.*, less adaptive iterations over the RIR tail portions), gives practically the same results for MDF and GMDF. Moreover, it turns out that the EW technique is essential for MDF with BIA, without which the overall performance decreases substantially, as it in effect applies the continuity constraint across individually adapted sub-blocks.

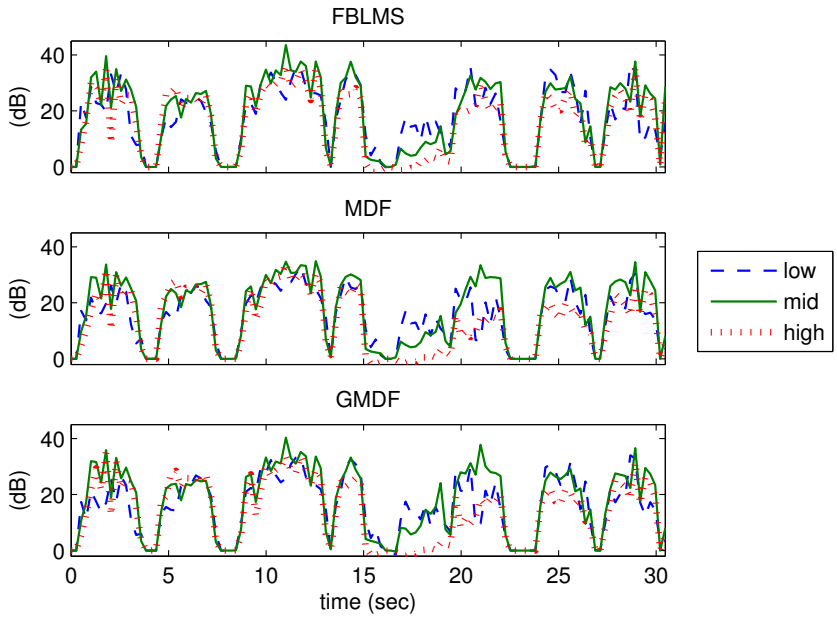


Figure 45: Sub-band tERLE decomposition (single channel).

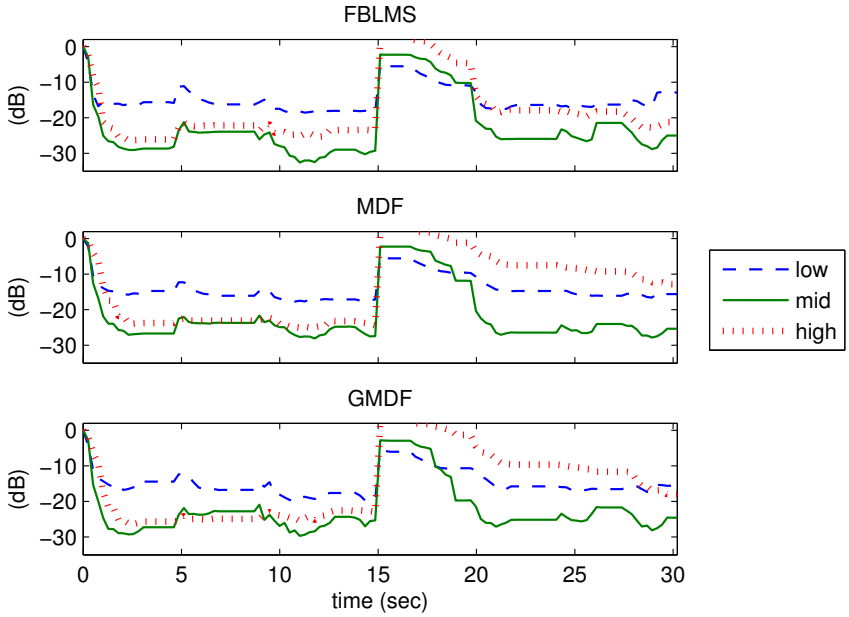


Figure 46: Sub-band misalignment decomposition (single channel).

Table 12: Average tERLE (dB, left column) and misalignment(dB, right column).

band	*FBLMS*		MDF		GMDF	
low	*20.6*	*-14.7*	17.9	-13.0	19.1	-14.7
mid	*22.2*	*-22.4*	20.9	-21.5	21.8	-22.1
high	*20.3*	*-17.7*	17.4	-13.6	18.7	-15.1
all	*21.5*	*-17.7*	18.7	-14.5	19.6	-15.7

CHAPTER V

ROBUST MCAEC VIA SBSS

The least mean square (LMS) algorithm [138] is summarized by the filter coefficient update equation

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu e(n)\mathbf{r}(n), \tag{102}$$

where $\mathbf{r}(n)$ is the reference signal vector, $\mathbf{w}(n)$ is the filter coefficient vector, μ is the adaptation step-size, and $e(n)$ is the estimation error. It relies on information from the second-order statistics (SOS) for adaptation as the mean-square error (MSE) $E\{e^2(n)\}$ is minimized to obtain the optimal filter coefficients that best represent the true echo path. There are two main issues that prevent the LMS algorithm from achieving a desired echo cancellation performance: the presence of local acoustic noise (or near-end speech) and the non-uniqueness problem, the latter of which arises during multi-channel AEC (MCAEC). To better handle the two problems, we propose a shift in the conventional MCAEC paradigm to a new framework of semi-blind source separation (SBSS). Blind source separation (BSS) is a powerful signal enhancement method for recovering a target signal from a mixture of signals when no prior information on the original source signals are available. SBSS is a direct extension of BSS when some partial knowledge of the source signals are already available (*e.g.*, components of the reference signals), and is naturally suited for the MCAEC purpose in the presence of multiple interfering signals.

BSS can be implemented in the frequency domain through a batch, *i.e.*, offline, adaptation based on independent component analysis (ICA), which aims to maximize the statistical independence of separated signal components given that the original sources themselves are independent. In particular, the natural gradient algorithm [3] has become a standard in realizing the ICA optimization and can also be interpreted as performing a nonlinear decorrelation by using the higher-order statistics (HOS) [56]. Besides computational efficiency and batch-wise adaptability, BSS is commonly carried out in the frequency domain since

a convolutive mixture becomes a linear and instantaneous mixture after the short-time Fourier transform (STFT), thus making the separation problem much more tractable than in the time domain. Although there are many problems associated with a practical implementation of the ICA-based BSS in the frequency domain, *e.g.*, the permutation and the scaling ambiguities, such a truly multi-channel approach to signal enhancement has been proven to be very effective in handling a convolutive mixture of speech signals. The SBSS framework was first proposed in [60] and was successfully implemented in [79] as a combination of multi-channel BSS and a single-channel AEC in the frequency domain. We have shown subsequently in [135] that BSS and stereophonic AEC (SAEC) can be effectively implemented together in such a framework and that a decorrelation procedure also helps SBSS achieve better echo cancellation performance just as in the MCAEC case.

In this chapter, as we have done in [89], we analyze the structure of the SBSS de-mixing matrix to see how the multi-channel echo cancellation performance can be improved. After a deep theoretical analysis, algorithmic issues are discussed to provide suggestions for the design of robust and practical SBSS systems. In particular, we examine the behavior of a combination of batch and online adaptations, which we refer to as a *batch-online* adaptation, to possibly take advantage of both types of learning. We ultimately show through different far-end mixing conditions that with a proper constraint on the de-mixing matrix and a regularization procedure, both high echo cancellation and relatively low misalignment can be achieved without any pre-decorrelation procedure even for the worst-case scenario of a single far-end talker along with the non-uniqueness condition on the far-end mixing system.

The rest of this chapter is organized as follows. First in Section 5.1, we present a deep theoretical analysis of the SBSS framework, including: presentation of the model of an SBSS system, discussion of the origin of the non-uniqueness problem in SBSS, steady-state analysis of the SBSS system, illustration of a connection between the MSE-based and the ICA-based approaches, and exploration of constraints on the SBSS de-mixing filter and their effect on the ICA optimization. Next in Section 5.2, we discuss the main issues of the algorithmic SBSS design, including: detailed outline of an online implementation of the SBSS system and description of the implemented SBSS algorithm. Finally in Section 5.3,

we describe methods for evaluating the SBSS performance and provide simulated and real-world results.

5.1 Generalization of MCAEC by SBSS

5.1.1 SBSS Model

We consider a time-invariant mixing model in the frequency domain that is as general as possible, and we assume zero-mean random processes that generate the involved signals. We will make some more simplifications later to make the analysis more tractable.

To begin with, we define the notations that are used in the following sections. A model for the near-end and the far-end mixing systems and the SBSS system is illustrated in Figure 47. At the far end, Q sources are recorded by an array of R microphones. At the near end, an array of S microphones records P near-end sources and R loudspeaker signals. That is, a far-end source signals vector $\mathbf{q}$ is multiplied by an $R \times Q$ frequency response matrix $\mathbf{G}$, which represents the far-end mixing system, to give the reference signal vector $\mathbf{r}$:

$$\mathbf{r}(\omega, t) = \mathbf{G}(\omega)\mathbf{q}(\omega, t), \tag{103}$$

where $\mathbf{r}(\omega,t) = [r_1(\omega,t), \cdots, r_R(\omega,t)]^T$ and $\mathbf{q}(\omega,t) = [q_1(\omega,t), \cdots, q_Q(\omega,t)]^T$. A near-end multi-channel source-signal vector $\mathbf{s}$ is then multiplied by an $S \times P$ frequency response matrix $\mathbf{H}_{11}$, and a multi-channel reference-signal vector $\mathbf{r}$ is multiplied by an $S \times R$ frequency response matrix $\mathbf{H}_{12}$, *i.e.*, the echo paths. The two matrices can be combined into a single $(S+R) \times (P+R)$ matrix $\mathbf{H}$ that represents the entire near-end mixing system:

$$\mathbf{x}(\omega, t) = \begin{bmatrix} \mathbf{x}_s(\omega, t) \\ \mathbf{x}_r(\omega, t) \end{bmatrix} = \mathbf{H}(\omega) \begin{bmatrix} \mathbf{s}(\omega, t) \\ \mathbf{r}(\omega, t) \end{bmatrix}, \tag{104}$$

where $\mathbf{s}(\omega,t) = [s_1(\omega,t), \cdots, s_P(\omega,t)]^T$, $\mathbf{x}_s(\omega,t) = [x_1(\omega,t), \cdots, x_S(\omega,t)]^T$, $\mathbf{x}_r(\omega,t) = [x_{S+1}(\omega,t), \cdots, x_{S+R}(\omega,t)]^T$, and

$$\mathbf{H}(\omega) = \begin{bmatrix} \mathbf{H}_{11}(\omega) & \mathbf{H}_{12}(\omega) \\ \mathbf{O}_{R\times P} & \mathbf{I}_R \end{bmatrix}_{(S+R)\times(P+R)}, \tag{105}$$

where $\mathbf{O}_{R\times P}$ is an $R \times P$ matrix with all elements equal to zero and $\mathbf{H}_{22}$ is automatically assigned to be an $R \times R$ identity matrix $\mathbf{I}_R$. Furthermore, by substituting (103) into (104),

the far-end and the near-end mixing systems can be combined into a unified mixing system represented by an $(S+R)\times(P+Q)$ matrix $\tilde{\mathbf{H}}$:

$$\mathbf{x}(\omega,t)=\tilde{\mathbf{H}}(\omega)\begin{bmatrix}\mathbf{s}(\omega,t)\\ \mathbf{q}(\omega,t)\end{bmatrix}, \tag{106}$$

$$\tilde{\mathbf{H}}(\omega)=\begin{bmatrix}\mathbf{H}_{11}(\omega) & \mathbf{H}_{12}(\omega)\mathbf{G}(\omega)\\ \mathbf{O}_{R\times P} & \mathbf{G}(\omega)\end{bmatrix}_{(S+R)\times(P+Q)}. \tag{107}$$

Hereafter, we assume $S=P$ and $R=Q$, which implies that the matrix $\tilde{\mathbf{H}}$ may be invertible. We will show later that with proper constraints on the adaptation, different conditions for the source numbers Q and P may occur for SBSS to be still effective for the MCAEC purpose.

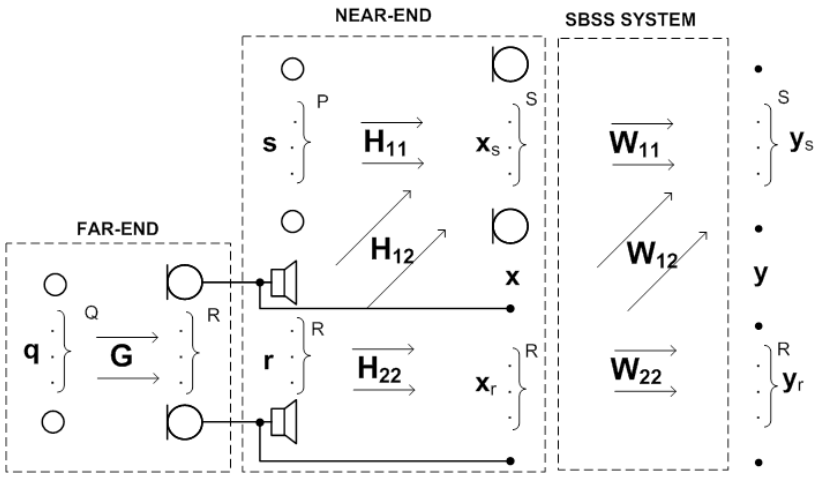


Figure 47: Model of the near-end and the far-end mixing systems and the semi-blind source separation (SBSS) system.

The goal of an SBSS system is to perform the estimation of the near-end source signals by using an $(S+R)\times(S+R)$ de-mixing matrix $\mathbf{W}$ such that

$$\mathbf{y}(\omega)=\begin{bmatrix}\mathbf{y}_s(\omega,t)\\ \mathbf{y}_r(\omega,t)\end{bmatrix}=\mathbf{W}(\omega)\mathbf{x}(\omega,t)\simeq\begin{bmatrix}\mathbf{s}(\omega,t)\\ \mathbf{q}(\omega,t)\end{bmatrix}, \tag{108}$$

where $\mathbf{y}_s(\omega,t)=[y_1(\omega,t),\cdots,y_S(\omega,t)]^T$ and $\mathbf{y}_r(\omega,t)=[y_{S+1}(\omega,t),\cdots,y_{S+R}(\omega,t)]^T$. We then generalize the structure of $\mathbf{W}$ as

$$\mathbf{W}(\omega)=\begin{bmatrix}\mathbf{W}_{11}(\omega) & \mathbf{W}_{12}(\omega)\\ \mathbf{O}_{R\times S} & \mathbf{W}_{22}(\omega)\end{bmatrix}_{(S+R)\times(S+R)}, \tag{109}$$

where $\mathbf{W}_{11} = [w_{ij}]_{1\leq i,j\leq S}$, $\mathbf{W}_{12} = [w_{ij}]_{1\leq i\leq S, S+1\leq j\leq R}$, and $\mathbf{W}_{22} = [w_{ij}]_{S+1\leq i,j\leq S+R}$. We can see from (106) and (108) that the optimal solution for $\mathbf{W}$ is obtained when $\mathbf{W}\tilde{\mathbf{H}} = \mathbf{I}$ such that

$$\mathbf{W}_{11}(\omega)\mathbf{H}_{11}(\omega) = \mathbf{I}_S, \tag{110}$$

$$\mathbf{W}_{22}(\omega)\mathbf{G}(\omega) = \mathbf{I}_R, \tag{111}$$

$$[\mathbf{W}_{11}(\omega)\mathbf{H}_{12}(\omega) + \mathbf{W}_{12}(\omega)]\mathbf{G}(\omega) = \mathbf{O}_{S\times R}. \tag{112}$$

That is, the SBSS system is able to jointly perform the separation of near-end source signals and the cancellation of acoustic echoes such that the diagonal terms become an identity while the off-diagonal terms become null. More specifically, $\mathbf{W}_{11}$ and $\mathbf{W}_{22}$ provide the separation of the near-end and the far-end sources, respectively, whereas $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$ are together responsible for the acoustic echo cancellation. There are two other key observations to be made from constraining the structure of $\mathbf{W}$ as in (109).

First, we must point out that we are not interested in recovering the signals played through the loudspeakers since we already have them as the reference signals. Moreover, we assume that the responsibility of separating the far-end microphone signals, *i.e.*, the reference signals, is with the far-end SBSS system, thus we do not have to adapt $\mathbf{W}_{22}$ for the purpose of obtaining the original source vector $\mathbf{q}$. Therefore, $\mathbf{y}_r$ can be any linear combination of $\mathbf{r}$, and the exact form of $\mathbf{W}_{22}$ may be controlled to optimize the near-end SBSS performance appropriately. The constraining of $\mathbf{W}_{22}$ and its effect on the SBSS adaptation process are discussed in more detail in Section 5.1.5.

The other observation is that $\mathbf{W}$ is a block upper-triangular matrix, *i.e.*, $\mathbf{W}_{21} = \mathbf{O}_{R\times S}$, since the reference signal channels are completely blind to the near-end source signals, *i.e.*, $\mathbf{H}_{21} = \mathbf{O}_{R\times S}$. However, the near-end signals themselves may take a round-trip in a full-duplex teleconferencing system, and the resulting echo may not be canceled entirely by the far-end SBSS system such that the reference signals would contain some remnants of the near-end signals. We neglect such a circumstance here by assuming that the far-end system has performed a reasonable job in suppressing the echo and that the latest instance of the near-end signals is uncorrelated with its own echo due to a sufficiently large round-trip delay.

Hence by requiring $\mathbf{W}_{21} = \mathbf{O}_{R\times S}$, we avoid the possibility that $\mathbf{y}_r$ contains the near-end source signals. In other words, the echo-canceled output components of the SBSS system in $\mathbf{y}_s$ are subject to the permutation ambiguity introduced only through the separation of the near-end sources, and the ambiguity problem then needs to be solved only for the sub-matrix $\mathbf{W}_{11}$.

5.1.2 Non-uniqueness Problem in SBSS

The non-uniqueness problem also occurs in SBSS and can be briefly analyzed as follows. If $\mathbf{W}_{11}$ and $\mathbf{G}$ are not singular, $\mathbf{H}_{12}$ that corresponds to the echo paths can be uniquely obtained from (112) as

$$\widehat{\mathbf{H}}_{12}(\omega) = -\mathbf{W}_{11}(\omega)^{-1}\mathbf{W}_{12}(\omega). \tag{113}$$

When there is an equal number of sources and microphones at the near end (*i.e.*, $P = S$), the physical interpretation of the frequency-domain BSS [102, 88] ensures that $\mathbf{W}_{11}$ is almost always non-singular by assuming spatial diversity and mutual independence between the sources. In the case of more sources than microphones (*i.e.*, the under-determined case $P > S$), $\mathbf{H}_{11}$ is not invertible, and there is no unique solution for $\mathbf{W}_{11}$. However, the estimate of $\mathbf{W}_{11}$ is not necessarily singular in such a case, and its inversion is still attainable for the estimation of $\mathbf{H}_{12}$. On the other hand, the severe ill-conditioning, or near-singularity, of $\mathbf{G}$ is a much more serious problem due to a sparse representation at each frequency. We have already mentioned that a unique identification of the echo paths by the MSE-based MCAEC is possible only if the far-end mixing matrix is not singular. Also, (112) indicates the dependence of the solution for $\mathbf{H}_{12}$ on $\mathbf{G}$ just as in the MCAEC case [118]. Therefore, by neglecting the rare occurrence of singularity of $\mathbf{W}_{11}$, the non-uniqueness problem in SBSS can be considered approximately equivalent to that of the traditional MCAEC system.

5.1.3 Derivation of Steady-State Solution

The global de-mixing matrix $\mathbf{W}$ can be estimated through the gradient-descent estimation procedure:

$$\mathbf{y}_n(\omega, t) = \mathbf{W}_n(\omega)\mathbf{x}(\omega, t), \tag{114}$$

$$\mathbf{W}_{n+1}(\omega) = \mathbf{W}_n(\omega) + \eta\mathbf{\Gamma}[\mathbf{W}_n(\omega), \mathbf{y}_n(\omega, t)], \tag{115}$$

where $\mathbf{\Gamma}$ is the updating term as a function of $\mathbf{W}_n$ and $\mathbf{y}_n$, η is the adaptation step-size, and n is the iteration index. $\mathbf{\Gamma}$ takes different forms according to the cost function that is to be minimized through the gradient-descent procedure. The separation of a mixture of non-stationary source signals can be achieved by minimizing either the second-order or the higher-order correlation among the output signals in $\mathbf{y}_n$. In a more general case of non-Gaussian source signals, the separation is possible through ICA by maximizing the mutual independence between the separated output components by using the HOS.

Any gradient-descent algorithm may be used for the estimation of $\mathbf{W}$. For the analysis of the steady-state solution, we consider an ICA optimization procedure based on the natural gradient algorithm and the cost function determined by the Kullback-Leibler divergence [6]:

$$D_{KL}[p(\mathbf{y})\|p(y_1), \ldots, p(y_{S+R})] = \int p(\mathbf{y}) \log \frac{p(\mathbf{y})}{\prod_{i=1}^{S+R} p(y_i)} \, d\mathbf{y} \tag{116}$$

where $p(\mathbf{y})$ is the joint probability density function (PDF) of the de-mixed output signals and $p(y_i)$ is the marginal PDF of each output signal. We also perform batch adaptation in order to have a sufficient amount of observation for statistically consistent estimation. According to the natural gradient algorithm, the gradient term is given by

$$\mathbf{\Gamma}[\mathbf{W}_n(\omega), \mathbf{y}_n(\omega, t)] = \left\{\mathbf{I}_{S+R} - E\{\Phi(\mathbf{y}_n(\omega, t))\mathbf{y}_n(\omega, t)^H\}\right\} \mathbf{W}_n(\omega), \tag{117}$$

where $\Phi(\cdot)$ is a nonlinear function, $\{\cdot\}^H$ is the Hermitian (conjugate) transpose operator, and $E\{\cdot\}$ is the expectation operator that, assuming stationarity, can be approximated by averaging over time. Then assuming that all of the near-end sources and the echo paths are mutually independent, (115) converges to a solution that minimizes the Kullback-Leibler divergence between the separated output signals given by (114).

The convergence analysis of the natural gradient algorithm is a very difficult task. Still, we can approximate the analysis through a direct inspection of what we refer to as the generalized covariance matrix $\mathrm{E}\{\Phi(\mathbf{y})\mathbf{y}^H\}$ by assuming that the components in the output vector $\mathbf{y}$ are zero-mean random variables such that they can be considered statistically independent when $\mathrm{E}\{\Phi(\mathbf{y})\mathbf{y}^H\}$ is diagonalized. The diagonalization is indeed achieved by minimizing each cross-moment of order $\mathtt{u}$ after a Taylor expansion of the nonlinear function $\Phi(\cdot)$:

$$E\{y_a^{\mathtt{u}}(\omega)y_b^*(\omega)\} = 0 \quad \forall \mathtt{u} \in \mathbb{N}, \tag{118}$$

where * denotes the complex conjugation. That is, the statistical independence for two output components y_a and y_b is achieved when the generalized covariance $E\{\Phi(y_a)y_b^*\}$ becomes zero, as two zero-mean random variables can be considered statistically independent if all of the higher-order cross-cumulants are zero [21].

We can analyze the structure of the steady-state solution for $\mathbf{H}_{12}$ as follows. First, the input signals for (114) are obtained by applying (105) to (104):

$$\begin{bmatrix} \mathbf{x}_s(\omega,t) \\ \mathbf{x}_r(\omega,t) \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{11}(\omega)\mathbf{s}(\omega,t) + \mathbf{H}_{12}(\omega)\mathbf{r}(\omega,t) \\ \mathbf{r}(\omega,t) \end{bmatrix}. \tag{119}$$

Next, the output signals used for updating $\mathbf{W}$ in (115) are obtained from (108) and (109):

$$\begin{bmatrix} \mathbf{y}_s(\omega,t) \\ \mathbf{y}_r(\omega,t) \end{bmatrix} = \begin{bmatrix} \mathbf{W}_{11}(\omega)\mathbf{H}_{11}(\omega)\mathbf{s}(\omega,t) + [\mathbf{W}_{11}(\omega)\mathbf{H}_{12}(\omega) + \mathbf{W}_{12}(\omega)]\,\mathbf{r}(\omega,t) \\ \mathbf{W}_{22}(\omega)\mathbf{r}(\omega,t) \end{bmatrix}. \tag{120}$$

Now, let's for the moment consider statistical independence between the separated sources vector $\mathbf{y}_s$ associated with the near-end system and the separated sources vector $\mathbf{y}_r$ associated with the reference signals. The optimal solution for $\mathbf{H}_{12}$ is obtained by setting $E\{\mathbf{y}_s^{\mathtt{u}}\mathbf{y}_r^H\} = \mathbf{O}_{S\times R}$ $\forall \mathtt{u} \in \mathbb{N}$. Then, after omitting the frequency and time dependencies for notation convenience, we obtain the following expression by applying (118) to (120):

$$E\{\mathbf{y}_s^{\mathtt{u}}\mathbf{y}_r^H\} = E\{[\mathbf{W}_{11}\mathbf{H}_{11}\mathbf{s} + (\mathbf{W}_{11}\mathbf{H}_{12} + \mathbf{W}_{12})\mathbf{r}]^{\mathtt{u}}\mathbf{r}^H\mathbf{W}_{22}^H\} = \mathbf{O}_{S\times R} \quad \forall \mathtt{u} \in \mathbb{N}, \tag{121}$$

where $\mathbf{y}_s^{\mathtt{u}}$ indicates the raising of each component of $\mathbf{y}_s$ to the integer power $\mathtt{u}$ (*i.e.*, the scalar sources y_a and y_b from (118) are simply substituted by the vectors $\mathbf{y}_s$ and $\mathbf{y}_r$). By

applying the binomial expansion, we can rewrite (121) as:

$$E\{[\mathbf{W}_{11}\mathbf{H}_{11}\mathbf{s}]^{\mathtt{u}}\mathbf{r}^H\mathbf{W}_{22}^H\} + E\{[(\mathbf{W}_{11}\mathbf{H}_{12}+\mathbf{W}_{12})\mathbf{r}]^{\mathtt{u}}\mathbf{r}^H\mathbf{W}_{22}^H\}+$$
$$E\{[\textstyle\sum_{k=1}^{\mathtt{u}-1}\frac{(\mathtt{u}-1)!}{k!(\mathtt{u}-1-k!)}(\mathbf{W}_{11}\mathbf{H}_{11}\mathbf{s})^{\mathtt{u}-1-k}\odot((\mathbf{W}_{11}\mathbf{H}_{12}+\mathbf{W}_{12})\mathbf{r})^k]\mathbf{r}^H\mathbf{W}_{22}^H\} = \mathbf{O}_{S\times R}\quad \forall \mathtt{u}\in\mathbb{N}, \tag{122}$$

where $\odot$ indicates is the Hadamard (element-wise) product. By using the multinomial expansion to further expand the additive terms with powers $\mathtt{u}$, $\mathtt{u}-1-k$, and k, it is possible to demonstrate that if $\mathbf{r}$ and $\mathbf{s}$ are statistically independent from each other, the first and the third terms in (122) are zero. In fact, all the matrix elements would be factorized as a sum of moments $E\{s_i^{\mathtt{u}}r_j\}$ that are zero for each $\mathtt{u}$ if s_i and r_j are zero-mean and mutually independent. It means the solution for $\mathbf{H}_{12}$ that satisfies (121) does not depend on the near-end sources, and the optimization is possible even though both the near-end and the far-end sources are active at the same time (*i.e.*, the double-talk situation). It follows then that we can substitute (103) into (122) to obtain

$$E\{[(\mathbf{W}_{11}\mathbf{H}_{12}+\mathbf{W}_{12})\mathbf{G}\mathbf{q}]^{\mathtt{u}}\mathbf{q}^H\mathbf{G}^H\mathbf{W}_{22}^H\} = \mathbf{O}_{S\times R}\quad \forall \mathtt{u}. \tag{123}$$

Since the far-end sources are assumed to be statistically independent, we can rewrite (123) as (see Appendix C)

$$[(\mathbf{W}_{11}\mathbf{H}_{12}+\mathbf{W}_{12})\mathbf{G}]^{\mathtt{u}}E\{\mathbf{q}^{\mathtt{u}}\mathbf{q}^H\}\mathbf{G}^H\mathbf{W}_{22}^H = \mathbf{O}_{S\times R}\quad \forall \mathtt{u}, \tag{124}$$

where $E\{\mathbf{q}^{\mathtt{u}}\mathbf{q}^H\}$ is the full-ranked generalized covariance matrix of the far-end source. If $\mathbf{W}_{22}$ and $\mathbf{G}$ are not singular, then (124) is satisfied when

$$\mathbf{W}_{11}\mathbf{H}_{12}+\mathbf{W}_{12} = \mathbf{O}_{S\times R}. \tag{125}$$

Finally, assuming that $\mathbf{W}_{11}$ is known and invertible, $\mathbf{H}_{12}$ can be estimated as

$$\widehat{\mathbf{H}}_{12} = -\mathbf{W}_{11}^{-1}\mathbf{W}_{12}, \tag{126}$$

which agrees with (113).

Several key observations can be made from the above derivation. First of all, going from (122) to (123) was made possible by assuming independence between the reference signals in $\mathbf{r}$ and the near-end source signals in $\mathbf{s}$. That is, the optimal solution for $\mathbf{H}_{12}$ that satisfies

(126) is ideally possible through the ICA optimization even though both the near-end and the far-ends sources are active at the same time. This is analogous to the fact that the MSE optimization during AEC is still possible even if there is a local noise as long as the reference signal and the noise are uncorrelated with each other [53, Chapter 6]. However, due to the noisy sample-wise estimate of the gradient of the MSE, the LMS algorithm may fail in identifying the true echo paths since its asymptotic performance depends on the adaptation step-size and echo-to-background power ratio (EBR) [53, Chapter 6]. Similarly, we must be careful how the gradient term in (117) is estimated in order to maintain the inherent ability of the natural gradient algorithm to converge to the optimal solution when there are multiple active near-end sources.

Second, we assume by (123) that there is always a solution $\mathbf{W}_{12} = -\mathbf{W}_{11}\mathbf{H}_{12}$ that maximizes the statistical independence of the output signals in $\mathbf{y}_s$, but the exact echo path identification is possible only if $\mathbf{W}_{11}$, $\mathbf{W}_{22}$, and $\mathbf{G}$ are fully ranked. During the derivation, we only considered the optimization by maximizing the mutual independence between the vectors $\mathbf{y}_s$ and $\mathbf{y}_r$. In a full optimization procedure, $\mathbf{W}_{11}$ and $\mathbf{W}_{22}$ should also be adapted to make the output components in $\mathbf{y}_s$ and $\mathbf{y}_r$ mutually independent from one another. According to (110) and (111), $\mathbf{W}_{11}$ and $\mathbf{W}_{22}$ are expected to be the inverses of $\mathbf{H}_{11}$ and $\mathbf{G}$, respectively, up to arbitrary permutation and scaling of rows. Also, as we already pointed out earlier, a physical configuration of the frequency-domain BSS system makes the singularity of both $\mathbf{W}_{11}$ and $\mathbf{W}_{22}$ a rare occurrence. Therefore, as we are not interested in separating the reference signals at the near end, we can focus on the structure of $\mathbf{W}_{22}$ such that the entire SBSS system can be made robust to ill-conditioned situations.

Third, due to the possibility of ill-conditioning of the far-end response matrix $\mathbf{G}$, the derivations from (123) to (126) may not be exact, and thus the optimal solution $\widehat{\mathbf{H}}_{12}$ may not be unique. Although the natural gradient algorithm through a batch adaptation would still converge to a solution for $\mathbf{W}_{12}$ that maximizes the output independence within the current observed data even if the non-uniqueness problem exists, the solution would also be affected by the changes in the far-end mixing system, and a continuous and stable sample-wise or batch-wise adaptation would not be possible. To avoid such a problem, traditional

MCAEC methods employ a decorrelation procedure to not only improve the convergence rate but also to keep the current estimate of the echo paths as close to the optimal solution as possible. Indeed, as demonstrated in [135], a decorrelation of the reference signals is sufficient to improve the conditioning of the generalized autocovariance matrix $E\{\Phi(\mathbf{r})\mathbf{r}^H\}$ such that the SBSS system can continue converging towards a unique solution. With regard to this observation, we should remember that the ICA optimization jointly uses the HOS from the observed signals, thus it should be less sensitive to the effect of the non-uniqueness problem so that a decorrelation procedure with less degree of distortion can be implemented when compared to the MSE-based MCAEC that uses only the SOS. Nevertheless, we show in Section 5.1.5 that a proper matrix constraint, which exploits the low spatial correlation between the far-end room impulse responses (RIRs) (observable under certain conditions [44]), is sufficient for reducing the fluctuations in the estimate of $\mathbf{H}_{12}$ during the ICA optimization so that the adaptation process becomes stable even without using any decorrelation procedure.

Finally, the joint adaptation of $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$ requires that the near-end source separation must also be applied. However, if simultaneous source separation and echo cancellation are not necessary, $\mathbf{W}_{12}$ can be re-scaled by multiplying by the inverse $\mathbf{W}_{11}^{-1}$ after adapting $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$ separately. The re-scaling, or normalization, process and its effect on the stability of SBSS are discussed further in Section 5.1.5.

5.1.4 Connection between MSE and ICA

Assuming that there are no active near-end sources, *i.e.*, $\mathbf{y}_s = \mathbf{0}_S$, where $\mathbf{0}_S$ is a zero vector of length S, the conventional MCAEC minimizes the MSE in each near-end microphone channel to obtain the optimal solution:

$$\begin{aligned}\widehat{\mathbf{h}}_i(\omega) &= \underset{\widetilde{\mathbf{h}}_j}{\operatorname{argmin}}\ E\{|e_i(\omega,t)|^2\} \\ &= \underset{\widetilde{\mathbf{h}}_i}{\operatorname{argmin}}\ E\{|[\mathbf{h}_i(\omega) - \widetilde{\mathbf{h}}_i(\omega)]^T\mathbf{r}(\omega,t)|^2\},\end{aligned} \tag{127}$$

where $e_i(\omega,t)$ is the estimation error for the i^{th} microphone channel, $1 \leq i \leq S$, $\mathbf{r} = [r_1, \cdots, r_R]^T$ is a vector of R loudspeaker (*i.e.*, reference) signals, and $\mathbf{h}_i = [h_{i1}, \cdots, h_{ij}, \cdots,$

$h_{iR}]^T$ is a vector of R frequency responses corresponding to the echo paths from the j^{th} loudspeaker to the i^{th} microphone. Taking the gradient of (127) with respect to $\widetilde{\mathbf{h}}_j$ and setting the result to zero gives

$$E\{e_i^*(\omega,t)\mathbf{r}(\omega,t)\} = E\{e_i(\omega,t)\mathbf{r}^*(\omega,t)\} = \mathbf{0}_R, \tag{128}$$

which is the well-known orthogonality principle, *i.e.*, the estimation error is decorrelated from the reference signals.

On the other hand, the echo paths determined by the ICA-based SBSS satisfies (123) for all integer powers $\mathtt{u}$. Since the near-end sources are assumed to be inactive, (123) can be simplified by setting $\mathbf{W}_{11} = \mathbf{I}_S$. Moreover, we can impose the constraint $\mathbf{W}_{22} = \mathbf{I}_R$ since we do not need to separate the reference signals at the near end. Substituting $\mathbf{r} = \mathbf{G}\mathbf{q}$ into the result gives

$$E\{[(\mathbf{H}_{12}(\omega) + \mathbf{W}_{12}(\omega))\mathbf{r}(\omega,t)]^{\mathtt{u}}\mathbf{r}^H(\omega,t)\} = \mathbf{O}_{S\times R} \ \ \forall \mathtt{u} \in \mathbb{N}. \tag{129}$$

Once an ICA optimization procedure converges to the steady-state solution $\mathbf{H}_{12} = -\mathbf{W}_{11}^{-1}\mathbf{W}_{12}$ $= -\mathbf{W}_{12}$, (129) can be rewritten for $\mathtt{u} = 1$ (*i.e.*, only the SOS are considered) as

$$\begin{aligned} E\{[\mathbf{H}_{12}(\omega) + \widehat{\mathbf{W}}_{12}(\omega)]\mathbf{r}(\omega,t)\mathbf{r}^H(\omega,t)\} = \\ E\{[\mathbf{H}_{12}(\omega) - \widehat{\mathbf{H}}_{12}(\omega)]\mathbf{r}(\omega,t)\mathbf{r}^H(\omega,t)\} = \\ E\{\mathbf{e}(\omega,t)\mathbf{r}^H(\omega,t)\} = \mathbf{O}_{S\times R}, \end{aligned} \tag{130}$$

where $\mathbf{e} = [e_1, \cdots, e_S]^T$ is a vector of estimation errors corresponding to S microphone signals. That is, the ICA-based SBSS minimizes the MSEs for all microphone channels such that the orthogonality principle is satisfied just as in the MSE-based MCAEC case.

The above result illustrates that the ICA-based SBSS is capable of *jointly* minimizing not only the MSE for every near-end microphone channel but also all the higher-order cross-correlations between the reference and the microphone channels through the diagonalization of the generalized covariance matrix $E\{\Phi(\mathbf{y})\mathbf{y}^H\}$ such that (129) is satisfied for all $\mathtt{u}$. The traditional MCAEC approach is inherently a single-channel AEC approach that simply minimizes the MSE at the output of each microphone independently from other microphones.

Thus the ICA-based SBSS should theoretically be able to handle multiple acoustic echoes better than the traditional MSE-based MCAEC since it has more information about the involved signals. Such a multi-channel adaptability based on the HOS is also the key to the ICA-based approach for being much more robust to the effects of the noise signals than the MSE-based approach that uses only the SOS, and it allows the ICA-based SBSS to perform simultaneous separation and cancellation of a mixture of multiple signals.

5.1.5 Constraint on Separation Matrix

We assume here that the uniqueness condition on the far-end mixing system is met such that the reference signals in $\mathbf{r}$ are linearly independent and that the generalized autocovariance matrix $E\{\Phi(\mathbf{r})\mathbf{r}^H\}$ is fully ranked. However, the actual conditioning of $E\{\Phi(\mathbf{r})\mathbf{r}^H\}$ varies largely according to the number of far-end sources and microphones due to sparse representations of the far-end signals and mixing system in the frequency domain.

To discuss the effects of constraining the global de-mixing matrix $\mathbf{W}$ on the SBSS adaptation process, we first assume that the number of the sources is equal to the number of microphones at the near end and consider three different mixing conditions at the far end:

1.(***Case A***: $Q = R$) The number of active sources is equal to the number of microphones.

2.(***Case B***: $Q > R$) The number of active sources is greater than the number of microphones.

3.(***Case C***: $Q < R$) The number of active sources is less than the number of microphones.

Afterward, we analyze the effect on the global adaptation for two different mixing conditions at the near end:

4.(***Case D***: $P = S$) The number of active sources is equal to the number of microphones.

5.(***Case E***: $P \neq S$) The number of active sources is different from the number of microphones.

5.1.5.1 *Case A: $Q = R$*

If the number of simultaneously active sources is equal to the number of microphones at the far end, the reference signals in $\mathbf{r}$ are guaranteed to be linearly independent and are correlated according to the impulse responses corresponding to an individual source. Although the near-end SBSS system is not responsible for the separation of the far-end source signals such that the estimation of $\mathbf{W}_{22}$ is not necessary, a progressive decorrelation of the reference signals during the adaptation would stabilize its convergence behavior since a stable minimum is approached when the entire generalized covariance matrix $E\{\Phi(\mathbf{y})\mathbf{y}^H\}$ is diagonalized.

Therefore, the first case to be examined involves no constraint on $\mathbf{W}_{22}$ to get a full benefit of the natural gradient algorithm based on ICA. However, the estimate of the matrix $\mathbf{W}$ may approach a singularity when no constraint is enforced on $\mathbf{W}_{22}$, thus we need to consider a normalization procedure for reducing the intrinsic scaling ambiguity of the ICA optimization. In particular, the scaling ambiguity for $\mathbf{W}$ can be reduced by applying the minimal distortion principle (MDP) [78]:

$$\widehat{\mathbf{W}}(\omega) = \mathrm{diag}[\mathbf{W}^{-1}(\omega)]\mathbf{W}(\omega). \tag{131}$$

The matrix $\mathbf{W}$ is not invertible if $\mathbf{W}_{22}$ is singular. However, when $\mathbf{W}$ has a block upper-triangular structure, its inverse is diagonalized:

$$\widehat{\mathbf{W}}(\omega) = \begin{bmatrix} \mathrm{diag}[\mathbf{W}_{11}^{-1}(\omega)] & \mathbf{O} \\ \mathbf{O} & \mathrm{diag}[\mathbf{W}_{22}^{-1}(\omega)] \end{bmatrix} \mathbf{W}(\omega). \tag{132}$$

Moreover, since we are not interested in the final output components corresponding to the decorrelated reference signals, we can avoid the inversion of the entire de-mixing matrix $\mathbf{W}$ and determine the scaling only for the near-end sources:

$$\widehat{\mathbf{W}}_{11}(\omega) = \mathrm{diag}[\mathbf{W}_{11}^{-1}(\omega)]\mathbf{W}_{11}(\omega). \tag{133}$$

In other words, the scaling ambiguity only depends on the separation of the near-end sources and not on the echo paths estimation. The idea is similar to limiting the permutation ambiguity in the estimate of $\mathbf{W}_{11}$ by enforcing $\mathbf{W}_{21} = \mathbf{O}$ as discussed in Section 5.1.1.

5.1.5.2 *Case B: $Q > R$*

If the number of active sources is greater than the number of microphones at the far end, the reference signals in $\mathbf{r}(\omega)$ are linearly independent and are decorrelated naturally by the presence of extra source signals beyond the number of microphones. Thus $E\{\Phi(\mathbf{y}_r)\mathbf{y}_r^H\}$ will be conditioned well, and the optimal solution for $\mathbf{H}_{12}$ can be found. However, $E\{\Phi(\mathbf{y}_r)\mathbf{y}_r^H\}$ can never be diagonalized since the number of observations is less than the number of sources. Consequently, an ICA optimization procedure may become unstable, and the iterative solution for $\mathbf{W}_{22}$ may never converge to a stable minimum. Hence, as the decorrelation of the reference signals through the adaptation of $\mathbf{W}_{22}$ is ill-advised in such a case, the overall stability of the system can be improved by constraining $\mathbf{W}_{22}$ to be a fixed matrix, *e.g.*, $\mathbf{W}_{22} = \mathbf{I}_R$.

5.1.5.3 *Case C: $Q < R$*

The optimal solution $\widehat{\mathbf{H}}_{12}$ may not be unique when there are fewer active sources than the microphones at the far end. Such a case corresponds to the near-singularity of the far-end response matrix $\mathbf{G}$ or equivalently to the rank deficiency of the generalized autocovariance matrix $E\{\Phi(\mathbf{r})\mathbf{r}^H\}$. In a real-life scenario, $E\{\Phi(\mathbf{r})\mathbf{r}^H\}$ should always be fully ranked since the modeling filters are generally shorter in length than the far-end RIRs. Also, nonlinear loudspeaker distortion and additive background noise naturally decorrelate the reference signals to help improve the conditioning of $E\{\Phi(\mathbf{r})\mathbf{r}^H\}$. However, the ill-conditioning of $\mathbf{G}$ in the frequency domain ultimately hampers an ICA optimization procedure from converging to the true echo paths, thus the overall echo cancellation performance becomes very sensitive to the variations in both the far-end and the near-end mixing systems. Nevertheless, although exploiting the HOS cannot solve the non-uniqueness problem, the likelihood that the gradient of the ICA optimization cost would point towards a specific region in the solution space during a gradient-descent adaptation is strongly related to the structure of the de-mixing matrix and to the characteristics of the far-end RIRs. For example, if the far-end microphones are sufficiently spaced apart as in a realistic situation (*e.g.*, 10 to 20

cm), the far-end RIRs are already sparse in the time domain. The sparseness is not necessarily inherited from the time domain at each frequency in the frequency domain, but the correlation between the frequency responses decreases with the microphones spacing if we assume the reverberation to be a diffuse noise field [44]. Then we can approximate the first factor of (124) as (after dropping ω for notational convenience)

$$[(\mathbf{W}_{11}\mathbf{H}_{12} + \mathbf{W}_{12})\mathbf{G}]^{\mathtt{u}} \simeq (\mathbf{W}_{11}\mathbf{H}_{12} + \mathbf{W}_{12})^{\mathtt{u}}\mathbf{G}^{\mathtt{u}}, \tag{134}$$

which can be derived as in Appendix C by considering $\mathbf{W}_{11}$, $\mathbf{H}_{12}$, and $\mathbf{W}_{12}$ to be constant matrices and $\mathbf{G}$ a matrix of zero-mean independent random variables, taking the expectation of $[(\mathbf{W}_{11}\mathbf{H}_{12} + \mathbf{W}_{12})\mathbf{G}]^{\mathtt{u}}$, and estimating $E\{\mathbf{G}^{\mathtt{u}}\}$ by $\mathbf{G}^{\mathtt{u}}$. Also, since the far-end sources are assumed to be independent, the generalized covariance matrix $E\{\mathbf{q}^{\mathtt{u}}\mathbf{q}^{H}\}$ is expected to be diagonal. Thus we can approximate (124) as

$$\mathbf{G}^{\mathtt{u}}E\{\mathbf{q}^{\mathtt{u}}\mathbf{q}^{H}\}\mathbf{G}^{H} \simeq \mathrm{diag}\{E\{q_i^{\mathtt{u}}q_i^{*}\}\sum_j g_{ij}^{\mathtt{u}}g_{ij}^{*}\} = \mathbf{D}, \tag{135}$$

where $\mathbf{D}$ is a diagonal matrix. Therefore, by constraining $\mathbf{W}_{22}$ to be an identity matrix $\mathbf{I}$, $\mathbf{W}$ takes the following structure:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{O} & \mathbf{I}_R \end{bmatrix}. \tag{136}$$

Assuming for simplicity $\mathbf{W}_{11} = \mathbf{I}_S$ (*i.e.*, no near-end source separation is performed) and using the $\mathbf{W}_{22} = \mathbf{I}_R$ constraint, (124) reduces to

$$E\{\mathbf{y}_s\mathbf{y}_r^{H}\} \simeq (\mathbf{H}_{12} + \mathbf{W}_{12})^{\mathtt{u}}\mathbf{D}. \tag{137}$$

It becomes clear from (137) that with the de-mixing matrix constraint of (136) and an "ideal" assumption of zero correlation between the far-end frequency responses (*i.e.*, infinite distance between the far-end microphones), the elements of the matrix $\mathbf{W}_{12}$ are independently optimized. In other words, the update direction during the gradient-descent optimization procedure for the $(i,j)^{th}$ element of $\mathbf{W}_{12}$ does not depend on the other elements that are related to different echo paths, and hence the effect of the non-uniqueness problem is alleviated through the reduction in the ambiguity of physically allowed solution of the echo paths.

We should point out that the diagonal constraint of $\mathbf{W}_{22}$ cannot completely solve the true non-uniqueness problem since (135) is only an approximation. Nonetheless, it ensures that the iterative solution for $\mathbf{W}_{12}$ is attracted more likely to a region very close to the point that corresponds to the true echo paths since the contribution of the off-diagonal terms in (135) can be neglected for many frequencies. Thus, the constraint tends to globally bind the solution space of the time-domain filters related to $\mathbf{W}_{12}$, which consequently reduces the overall misalignment.

5.1.5.4 ***Case D**: $P = S$*

If the number of active sources is equal to the number of microphones at the near end, there exists an invertible de-mixing matrix $\mathbf{W}_{11}$ that maximize the output independence of the estimated near-end sources $\mathbf{y}_s$. Therefore, the estimation of the echo paths is possible by applying (126).

5.1.5.5 ***Case E**: $P \neq S$*

If the number of active sources is different from the number of microphones at the near end, we can examine two sub-cases. For the case of $P < S$, the mixing system is not fully ranked, and it may not be possible to estimate an invertible de-mixing matrix that corresponds to the inverse of $\mathbf{H}_{11}$. However, since the natural gradient algorithm does not apply any matrix inversion, the adaptation may still converge to a singular matrix that maximizes the output independence of the estimated near-end sources $\mathbf{y}_s$ (in such a case, some rows of $\mathbf{W}_{11}$ would be zero). For the case of $P > S$, since the matrix $\mathbf{H}_{11}$ is not square, the natural gradient algorithm cannot converge to an $S \times S$ de-mixing matrix capable of separating the near-end source mixture and may approach singularity during the adaptation of $\mathbf{W}_{11}$.

From the MCAEC perspective, the singularity of $\mathbf{W}_{11}$ should be avoided for both of the sub-cases $P < S$ and $P > S$ since it would hamper the estimation of the true echo paths by (126) and of the output sources by the MDP of (133). Therefore, constraints for preventing the singularity of $\mathbf{W}_{11}$ are required. For example, if the near-end source separation is not the main goal of the SBSS system, the constraint $\mathbf{W}_{11} = \mathbf{I}_S$ may be adopted. Alternatively, the singularity of $\mathbf{W}$ can be avoided by regularized variants of the natural gradient algorithm

such as the Flexible ICA proposed in [22]. However, it is worth underlining that in real-world scenarios, the singularity is a very rare occurrence since the number of independent acoustic sources is in general greater than the number of the microphones. Furthermore, we note that the loudspeakers themselves represent independent sources at the near end. Hence the singularity of $\mathbf{W}$ may theoretically occur only when both the near-end sources and the acoustic echoes are absent, which is a degenerate case controlled easily by inhibiting the adaptation when the reference signals are zero.

Thus similarly to the permutation and scaling ambiguities, it may be concluded that the ambiguity of the source number needs to be solved only if the separation of the near-end sources is desired. On the other hand, the MCAEC is not affected by such an ambiguity since the number of independent sources that generate the echoes (*i.e.*, the loudspeakers) is constrained most often to be equal to the number of reference signals, which is fixed in advanced by the structure of $\mathbf{W}$ with the constraint $\mathbf{W}_{22} = \mathbf{I}_R$.

5.2 Algorithm Design and Related Issues

5.2.1 Online Implementation of SBSS

A batch implementation of the SBSS system was considered in the previous discussions. However, online structures need to be analyzed for a practical scenario. We note that the adaptation is still performed in the frequency domain and that the term "online" does not necessarily refer to the "sample-online" adaptation performed by a traditional time-domain AEC procedure.

We identify two main possible structures that need to be examined:

1. Online adaptation.

2. Batch-online adaptation.

5.2.1.1 Online Adaptation

The online estimation of $\mathbf{W}$ can be performed by considering the instantaneous evaluation of the frequency-domain values of $\mathbf{y}$. By substituting the iteration index n with the current time index t, and the expectation in (117) with the instantaneous generalized covariance

matrix $\Phi(\mathbf{y})\mathbf{y}^H$, the ICA solution is updated with the incoming data by iterating over the following formulas:

$$\mathbf{y}(\omega,t) = \mathbf{W}_t(\omega)\mathbf{x}(\omega,t), \tag{138}$$

$$\mathbf{W}_{t+1}(\omega) = \mathbf{W}_t(\omega) + \eta\{\mathbf{I}_{S+R} - \Phi[\mathbf{y}(\omega,t)]\mathbf{y}(\omega,t)^H\}\mathbf{W}_t(\omega), \tag{139}$$

where η is the adaptation step-size of the online adaptation. The choice of η plays an important role in the overall stability of the adaptation process. For example, a large value is preferred for quick adaptation to the changes in the mixing conditions, but alternatively a small value increases the accuracy of the steady-state solution. By using a fixed step-size, the stability of an online adaptation can be compromised by abrupt variations in the system, and (139) may easily diverge.

Among several solutions, a promising method for stabilizing the convergence behavior of the natural gradient algorithm consists of using the *a posteriori* unit-norm constraint on $\Phi(\mathbf{y})\mathbf{y}^H$. Such a normalized version of the natural gradient is referred to as the scaled natural gradient [27]. The need for scaling through normalization becomes even more relevant when the constraint $\mathbf{W}_{22} = \mathbf{I}_R$ is enforced. In fact, $\mathbf{W}$ is re-scaled by the intrinsic normalization effect of the natural gradient algorithm that regularizes the convergence behavior and ensures the convergence of the generalized covariance matrix, *i.e.*,

$$\Phi[\mathbf{y}(\omega,t)]\mathbf{y}(\omega,t)^H \rightarrow \mathbf{I}_{S+R}, \tag{140}$$

which is tightly linked to the estimation of the HOS, the accuracy of which is often limited by the lack in the amount of available data. However, if the $\mathbf{W}_{22} = \mathbf{I}_R$ constraint is enforced, the normalization effect does not apply to $\mathbf{W}_{22}$, and hence the norm of the de-mixing matrix may increase and possibly lead to divergence. Furthermore, even without any constraint on $\mathbf{W}_{22}$, the norm may increase also when the power of one of the reference signals becomes very small.

A common approach for stabilizing the adaptation process is to remove the normalization effect by modifying the structure of the natural gradient with a non-holonomic constraint:

$$\mathbf{W}_{t+1}(\omega) = \mathbf{W}_t(\omega) + \eta\{\mathbf{\Lambda_t}(\omega) - \Phi[\mathbf{y}(\omega,t)]\mathbf{y}(\omega,t)^H\}\mathbf{W}_t(\omega), \tag{141}$$

where $\mathbf{\Lambda}_t$ is a diagonal matrix. The adaptation becomes stable by using $\mathbf{\Lambda}_t = \text{diag}[\Phi(\mathbf{y})\mathbf{y}^H]$ even when the power of the signals changes rapidly over time [4]. However, the non-holonomic constraint does not guarantee the diagonality of $\mathbf{W}_{22}$, which in effect would increase the misalignment with a subsequent degradation in the overall echo cancellation performance according to the analysis in Section 5.1.5. Therefore, the use of the scaled natural gradient is preferable since the stabilization can be obtained while still preserving the constrained structure of the matrix $\mathbf{W}_{22}$.

We note that other diagonal constraints on $\mathbf{W}_{22}$ are possible, but we focus in this work on the applicability of the efficient scaling normalization. Specifically, we impose the scaled natural gradient along with the following combined constraints:

$$\mathbf{W}_{22} = \mathbf{I}_R, \;\; \Delta\mathbf{W}_{21} = \mathbf{O}_{R\times S}, \;\; \Delta\mathbf{W}_{22} = \mathbf{O}_{R\times R}, \tag{142}$$

where $\Delta\mathbf{W}_{21}$ and $\Delta\mathbf{W}_{22}$ are the sub-matrices relative to the gradient

$$\begin{aligned}\Delta\mathbf{W}_t(\omega) &= \left\{\mathbf{I} - \frac{1}{d(\omega,t)}\Phi[\mathbf{y}(\omega,t)]\mathbf{y}(\omega,t)^H\right\}\mathbf{W}_t(\omega) \\ &= \begin{bmatrix} \Delta\mathbf{W}_{11}(\omega) & \Delta\mathbf{W}_{12}(\omega) \\ \Delta\mathbf{W}_{21}(\omega) & \Delta\mathbf{W}_{22}(\omega) \end{bmatrix},\end{aligned} \tag{143}$$

where d is an inverse scaling factor. After the calculation of (143) with the constraints in (142), the matrix $\mathbf{W}$ is updated by

$$\mathbf{W}_{t+1}(\omega) = c(\omega,t)[\mathbf{W}_t(\omega) + \eta\Delta\mathbf{W}_t(\omega)], \tag{144}$$

where c is the scaling normalization. The scaling factors c and d are computed as in [27].

An online optimization is preferred for its ability in tracking the changes in the mixing conditions and for small input-output algorithm processing delay. However, the statistical bias in the instantaneous generalized covariance matrix cannot be neglected, especially when the adaptation is controlled through the HOS. Furthermore, the scaling and the permutation ambiguities of $\mathbf{W}_{11}$ cannot be easily solved through the instantaneous observations. To address these drawbacks, a batch-online approach is examined next.

5.2.1.2 Batch-Online Adaptation

The statistical bias in the instantaneous covariance matrix can be reduced by adopting a batch-online adaptation, where the higher-order correlations between the de-mixed output signal components is estimated over a certain number of time observations. Accordingly, (138) and (139) are then modified as

$$\mathbf{y}_n^{(b)}(\omega,t) = \mathbf{W}_n(\omega)\mathbf{x}^{(b)}(\omega,t), \tag{145}$$

$$\mathbf{W}_{n+1}(\omega) = \mathbf{W}_n(\omega) + \eta\{\mathbf{I}_{S+R} - E\{\Phi(\mathbf{y}_n^{(b)}(\omega,t))\mathbf{y}_n^{(b)}(\omega,t)^H\}\}\mathbf{W}_n(\omega) \tag{146}$$

where $\mathbf{x}^{(b)}$ is a vector of input signals in the b^{th} batch and $\mathbf{y}_n^{(b)}$ is a vector of output signals obtained at the n^{th} iteration in the b^{th} batch. Then in a batch-online implementation, the solution for $\mathbf{W}$ is recursively refined for a certain number of iterations by using the expected generalized covariance matrix $E\{\Phi(\mathbf{y}_n^{(b)})(\mathbf{y}_n^{(b)})^H\}$.

The expectation operator $E\{\cdot\}$ may be approximated by averaging over time in a batch adaptation procedure. On the other hand, it is estimated through a moving average procedure in a batch-online approach:

$$\begin{aligned} &E\{\Phi[\mathbf{y}_n^{(b)}(\omega,t)]\mathbf{y}_n^{(b)}(\omega,t)^H] = \\ &\quad \mu E\{\Phi[\mathbf{y}^{(b-1)}(\omega,t)]\mathbf{y}^{(b-1)}(\omega,t)^H\} + (1-\mu)A\{\Phi[\mathbf{y}_n^{(b)}(\omega,t)]\mathbf{y}_n^{(b)}(\omega,t)^H\}, \end{aligned} \tag{147}$$

where μ is a smoothing parameter that controls the averaging across batches, $E\{\Phi[\mathbf{y}^{(b-1)}]$ $(\mathbf{y}^{(b-1)})^H\}$ is the generalized covariance matrix estimated from the previous batch, and $A\{\cdot\}$ is the time-averaging operator defined as:

$$A\{\Phi[\mathbf{y}_n^{(b)}(\omega,t)]\mathbf{y}_n^{(b)}(\omega,t)^H\} = \frac{1}{T_{b+1}-T_b}\int_{T_b}^{T_{b+1}} \Phi[\mathbf{y}_n^{(b)}(\omega,t)]\mathbf{y}_n^{(b)}(\omega,t)^H dt, \tag{148}$$

where T_b and T_{b+1} defines the time interval of the data used in the b^{th} batch. We note that the approximation of the expectation by a moving average model is valid only if we assume ergodicity and stationarity of the random processes. However, the full ergodicity and stationarity cannot always be guaranteed, especially when the ICA cost function involving the HOS is used since the source statistics is almost always expected to evolve over time. In other words, the higher-order correlations cannot be estimated accurately by averaging over

time, which would result in an instability of the recursively estimated $E\{\Phi[\mathbf{y}_n^{(b)}](\mathbf{y}_n^{(b)})^H\}$, which in turn translates to fluctuations in the estimated $\mathbf{W}$.

To avoid the drawbacks of a recursive estimation of $E\{\Phi[\mathbf{y}_n^{(b)}](\mathbf{y}_n^{(b)})^H\}$, an alternative approach proposed in [85] is used that consists of the approximation

$$E\{\Phi[\mathbf{y}_n^{(b)}(\omega,t)]\mathbf{y}_n^{(b)}(\omega,t)^H\} \simeq A\{\Phi[\mathbf{y}_n^{(b)}(\omega,t)]\mathbf{y}_n^{(b)}(\omega,t)^H\}, \tag{149}$$

where the output signal components in $\mathbf{y}_n(\omega,t)$ are obtained by initializing the iteration in (145) and (146) with the matrix obtained at the $(b-1)^{th}$ batch. It means that we avoid the estimation of the generalized covariance matrix that depends on the HOS, and we enforce an online-type linking between batches through a proper matrix initialization.

5.2.2 Proposed SBSS Algorithm

We describe here a practical implementation of a batch-online SBSS algorithm described in Section 5.2.1 to verify the effectiveness of the constraints discussed in Section 5.1.5.

The time-domain microphone and reference signals in $\mathbf{x}(t) = [\mathbf{x}_s^T(t), \mathbf{x}_r^T(t)]^T$ are transformed through the STFT

$$\mathbf{x}(k,l) = \mathrm{STFT}[\mathbf{x}(t)], \tag{150}$$

where k is the frequency bin index and l is the block index in time. Signals are divided into non-overlapping batches of STFT blocks. The estimate of the de-mixing matrix $\mathbf{W}(k)$ for each frequency bin is adapted in each batch by recursing over the algorithm summarized in Table 13 for n_{max} iterations. The adaptation of the de-mixing matrix across batches is summarized in a pseudo-code in Table 14.

We note that the permutation ambiguity problem for $\mathbf{W}_{11}$ still exists and needs to be solved also by the SBSS algorithm. Permutation ambiguity is currently an open problem for ICA-based frequency-domain BSS, and many methods have been proposed. For example, TDOA-based approaches are reliable and robust in a batch adaptation even when the observed signals are very short in time [88, 101], where they are expected to be just as effective in a batch-online SBSS.

Table 13: Outline of the proposed SBSS algorithm.

1. Generalized covariance matrix estimation
$$\mathbf{y}(k,l) = \mathbf{W}(k)\mathbf{x}(k,l)$$
$$E\{\Phi[\mathbf{y}(k,l)]\mathbf{y}(k,l)^H\} = A\{\Phi[\mathbf{y}(k,l)]\mathbf{y}(k,l)^H\}$$
Computation of the scaling factors $c(k)$ and $d(k)$ according to [27]
2. Adaptation with the $\mathbf{W}_{22} = \mathbf{I}$ constraint
$$\Delta\mathbf{W}(k) = \{\mathbf{I}_{S+R} - \tfrac{1}{d(k)}E\{\Phi(\mathbf{y}(k,l))\mathbf{y}(k,l)^H\}\}\mathbf{W}(k)$$
$$\mathbf{W}_{22} = \mathbf{I}_R,\ \Delta\mathbf{W}_{21} = \mathbf{O}_{R\times S},\ \Delta\mathbf{W}_{22} = \mathbf{O}_{R\times R}$$
$$\mathbf{W}(k) = c(k)[\mathbf{W}(k) + \eta\Delta\mathbf{W}(k)]$$

Table 14: Procedure for adaptation of the de-mixing matrix.

$\widehat{\mathbf{W}}(k) = \mathbf{I}_{S+R}$;
while *b*
  for *k=1 to* N_k
    $\mathbf{W}(k) = \widehat{\mathbf{W}}(k)$
    for *n=1 to* n_{max}
      refine $\mathbf{W}(k)$ *as in Table 13*
    end for
    $\widehat{\mathbf{W}}(k) = \mathbf{W}(k)$
    $\mathbf{W}_{11}(k) = \mathrm{diag}(\mathbf{W}_{11}^{-1}(k))\mathbf{W}_{11}(k)$
    Solve permutation for $\mathbf{W}_{11}(k)$
  end for
end while

5.3 Experimental Evaluation

5.3.1 Evaluation Methods

We discuss here some important issues that need to be taken into account for a fair performance evaluation of an SBSS system.

For the traditional MSE-based AEC, the identification of the echo path is performed by minimizing the energy of the estimation error

$$e(t) = x(t) - \widehat{h}(t) * r(t) = x(t) - \sum_{\tau=0}^{L-1} \widehat{h}(\tau) r(t-\tau), \tag{151}$$

where L is the length of estimated filter $\widehat{h}$ in the time domain. Since the estimation error at time t is obtained by using the previous L taps from the reference signal r during the adaptation process, $\widehat{h}$ is constrained to be causal. On the other hand, in the batch-online SBSS, the de-mixing matrix $\mathbf{W}$ is estimated by evaluating the generalized covariance of the de-mixed signals over time observations within a batch of STFT blocks. Consequently, the estimated de-mixing matrix does not generally correspond to causal filters. Therefore, for a correct measurement of the misalignment, we need to transform the obtained filters from a non-causal to a causal one by a circular rotation of $L/2$ taps:

$$\widehat{\mathbf{H}}(t) = IFFT[-\mathbf{W}_{11}^{-1}(\omega)\mathbf{W}_{12}(\omega)], \tag{152}$$

$$\widehat{\mathbf{H}}_{causal}(t) = \text{shift}[\widehat{\mathbf{H}}(t), L/2], \tag{153}$$

where $\widehat{\mathbf{H}}$ represents a set of time-domain filters corresponding to all possible echo paths.

The causality of the estimated filters depend strictly on the rank of the far-end response matrix $\mathbf{G}$. If the rank of $\mathbf{G}$ is full, the optimal solution for the echo paths can be obtained regardless of the adaptation technique. In such a case, the time-domain filters in $\widehat{\mathbf{H}}$ are already causal, and a circular shifting would introduce only a delay without any effect on the echo cancellation. However, when the rank of $\mathbf{G}$ is not full, the ICA optimization procedure can converge to a solution for $\mathbf{W}_{12}$ that is not unique and whose physical interpretation does not match with the true echo paths represented by $\mathbf{H}_{12}$, in which case the corresponding time-domain filters are not causal.

To measure the echo cancellation performance, the following two metrics are used. First, the misalignment is defined by

$$\text{Misalignment}(i,j) \equiv 10\ \log_{10} \frac{||\mathbf{h}_{ij}(t) - \widehat{\mathbf{h}}_{ij}(t)||^2}{||\mathbf{h}_{ij}(t)||^2}, \tag{154}$$

where $\mathbf{h}_{ij}$ is a RIR vector corresponding to the echo path from the j^{th} loudspeaker to the i^{th} microphone and $\widehat{\mathbf{h}}_{ij}$ is the estimated RIR vector. Second, the *true* echo return loss enhancement (tERLE) is defined as

$$\text{tERLE}(i) \equiv 10\ \log_{10} \frac{|x_i(t) - \sum_j \tilde{\mathbf{h}}_{ij}^T(t)\mathbf{s}_j(t)|^2}{|y_i(t) - \sum_j \tilde{\mathbf{h}}_{ij}^T(t)\mathbf{s}_j(t)|^2}, \tag{155}$$

where $\tilde{\mathbf{h}}_{ij}$ is a vector corresponding to the RIR from the j^{th} near-end source to the i^{th} microphone. (155) is simply the traditional ERLE calculated after removing the near-end source signal so that the true amount of echo cancellation can be calculated during noisy time.

It is important to stress that for a fair evaluation of the SBSS performance, the above two metrics must be used with extra caution. The misalignment is an objective measure of the system identification performance and is in general indicative of the actual echo cancellation performance. However, large misalignment does not necessarily correspond to poor echo cancellation performance since a high degree of echo cancellation can still be achieved even with non-causal filters whose physical interpretation is not related to the true system identification, in which case the tERLE is the only meaningful metric for the performance evaluation.

5.3.2 Experimental Results

The SBSS algorithm proposed in Section 5.2.2 was evaluated on both simulated and real-world data. First, to better control the environmental conditions and to evaluate the effect of the parameters, we simulated the case of two pairs of near-end sources and microphones (*i.e.*, $P = 2$ and $S = 2$) and two pairs of far-end sources and microphones (*i.e.*, $Q = 2$ and $R = 2$). From the misalignment perspective, the theoretically worst-case scenario of single far-end talker and rapid changes in the far-end RIRs were simulated by alternating the activity of two far-end sources every 25 seconds as illustrated in Figure 48.

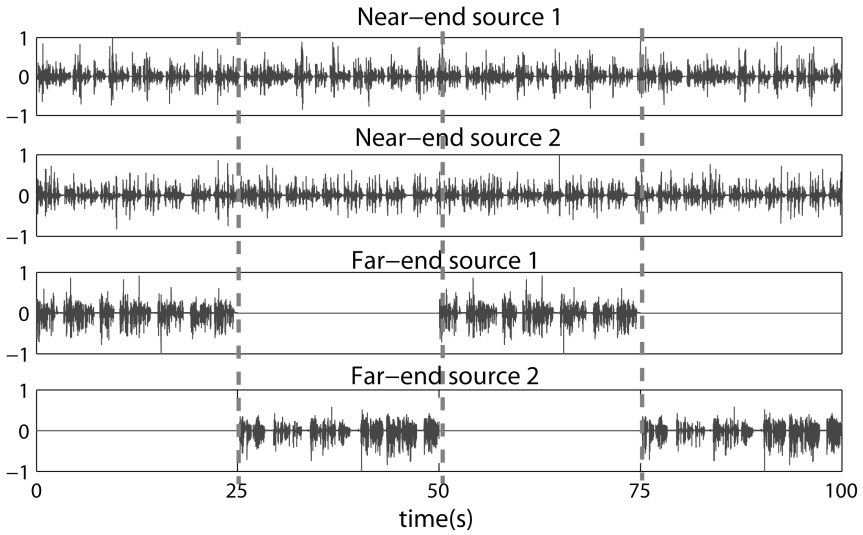


Figure 48: Source activities in the worst-case scenario.

A Hanning window of 4096 taps with 75% overlap was applied to speech signals sampled at $f_s = 16$ kHz before taking the STFT. The RIRs were simulated with by the Lehmann & Johansson's image source method [69]. The far-end and the near-end RIRs were truncated to 4096 and 3200 taps, respectively. The non-uniqueness problem should then be expected since $L = M = 4096$. The batch-online adaptation was evaluated with non-overlapping batches, where each batch lasted for one second to avoid more than one far-end source being active within a same batch and maintain the worst-case scenario. The algorithm parameters used during the adaptation are summarized in Table 15.

Table 15: Summary of parameters used during adaptation.

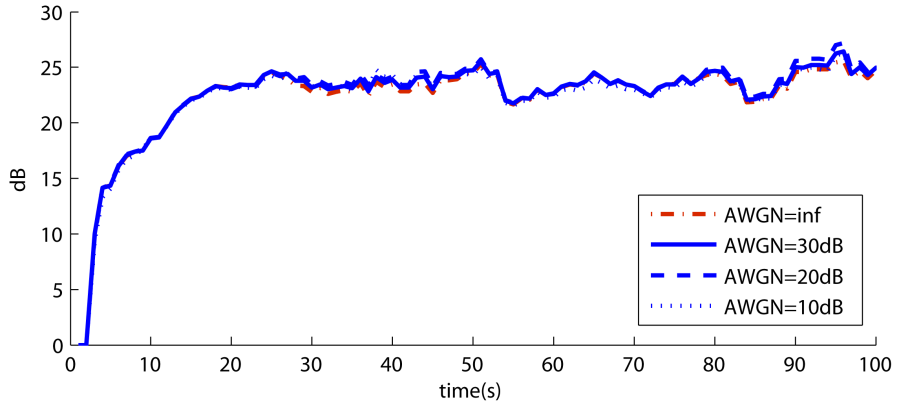
Parameters
$\eta = 0.1$
$\Phi(x) = \tanh(10 \cdot \lvert x \rvert) \exp(j\phi(x))$
$1 \leq n_{max} \leq 20$ (depending on the test situation)

For a fair performance evaluation, we considered an implementation of the batch-online adaptation without any *input-output* delay. That is, the data in the b^{th} batch are processed with a filter estimated from the previous $(b-2)^{th}$ batch. Such a method takes into account both the algorithmic and the computational delays assuming an unitary real-time factor.

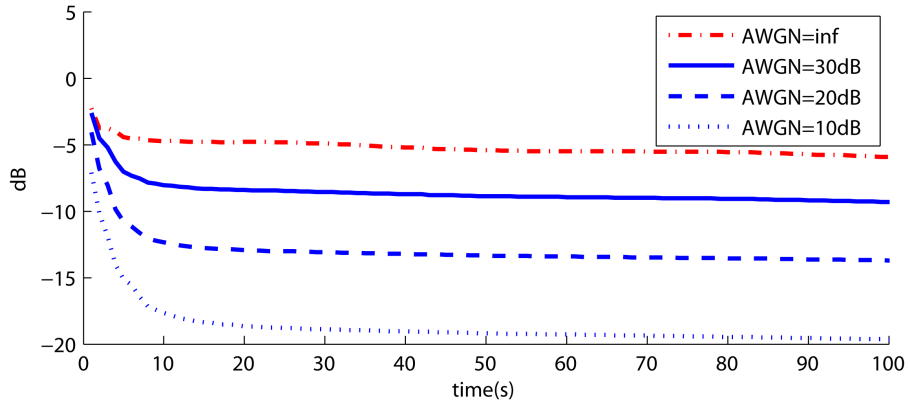
Thus the tERLEs for first two batches at the start of the adaptation are always equal to 0 dB. The tERLE and the misalignment are averaged over all possible combinations of signals and echo paths, respectively, to obtain a single performance measure for the entire system. Furthermore, the tERLE measurements have been smoothed over time with a first-order autoregressive moving average, with smoothing parameter equal to 0.8, for a more clear graphical representation.

Figure 49 shows the SBSS performance when additive white Gaussian noise (AWGN) is used to decorrelate the reference signals, where "AWGN=inf" corresponds to the case of no decorrelation procedure as the decorrelation level is measured in terms of the signal-to-noise ratio (SNR). We note that AWGN is used here as a practical choice to compare the SBSS performance with and without a decorrelation procedure and is not representative of the best procedure. As expected, the misalignment is reduced as the SNR is decreased to better decorrelate the reference signals. The misalignment converges to a relatively low value regardless of the presence of the near-end source signal, which evidently indicates the double-talk robustness of the SBSS algorithm. However, we note that the tERLE is almost equivalent for any value of SNR. As we have already stated previously, the optimal solution obtained through SBSS is not always equivalent to the actual echo paths, and an effective echo cancellation is possible even if the reference signals are linearly dependent.

Figure 50 shows the performance of SBSS with and without the $\mathbf{W}_{22} = \mathbf{I}_R$ constraint. We observe that for the constrained SBSS, the tERLE rapidly converges to a value of about 23 dB while the misalignment slowly decreases. As the two far-end sources alternate in activity every 25 seconds, the unconstrained SBSS displays a large degradation in the tERLE since the estimation of $\mathbf{W}_{12}$ depends on the far-end mixing condition. A better understanding can be obtained from the behavior of the unconstrained SBSS during the first 25 seconds, where the misalignment is considerably high even though the echo cancellation performance is acceptable. Again, it means that the SBSS algorithm converges to a solution that is strongly dependent on the far-end mixing system, and since a non-causal filtering was used, the solution does not have to make a physical sense. Such a solution is then valid for only one of the far-end source signals with which the adaptation was performed and



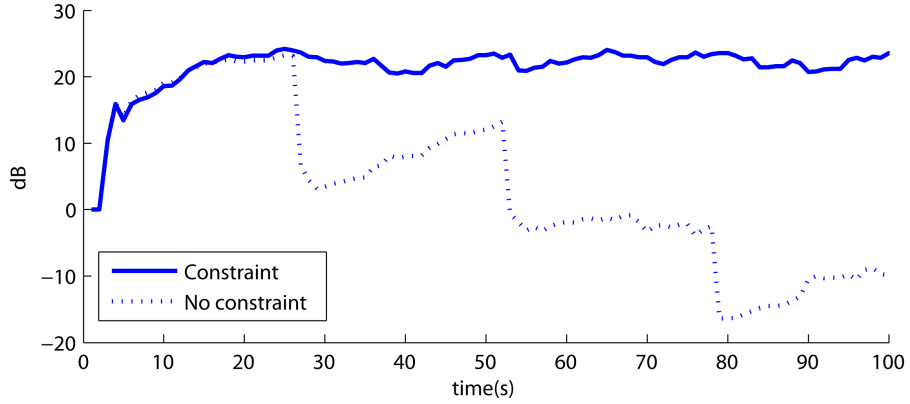
(a) True ERLE.



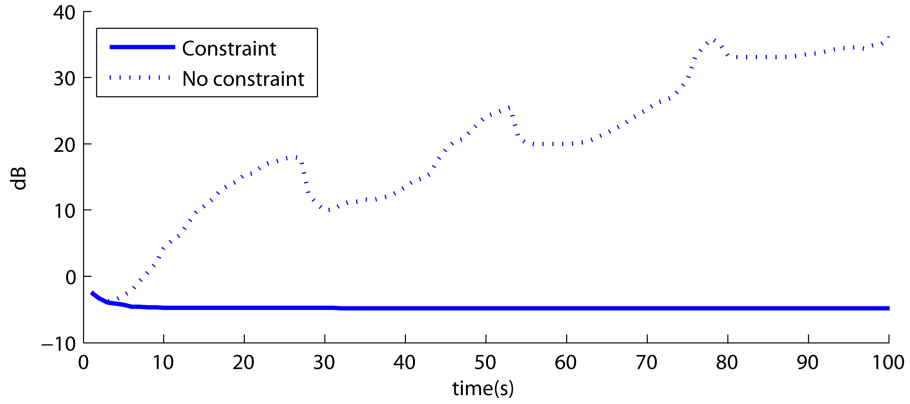
(b) Misalignment.

Figure 49: Performance of SBSS with AWGN decorrelation procedure (with de-mixing matrix constraints and 5 iterations).

not for the other, which explains a large degradation in tERLE for the unconstrained SBSS when the far-end source activity changes.



(a) True ERLE.

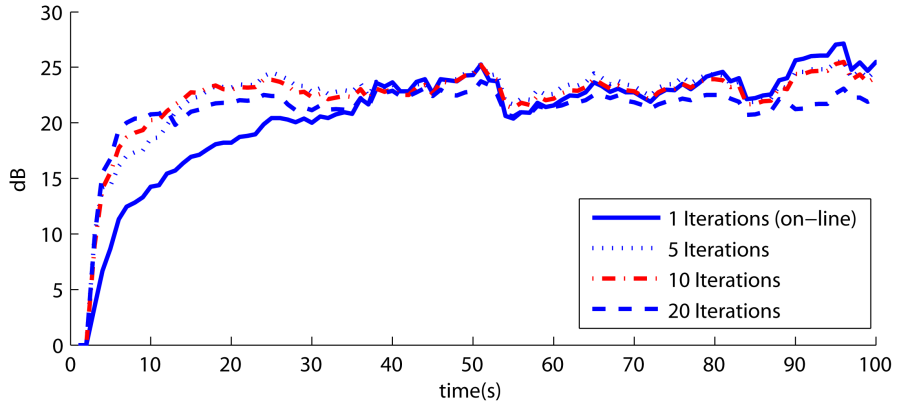


(b) Misalignment.

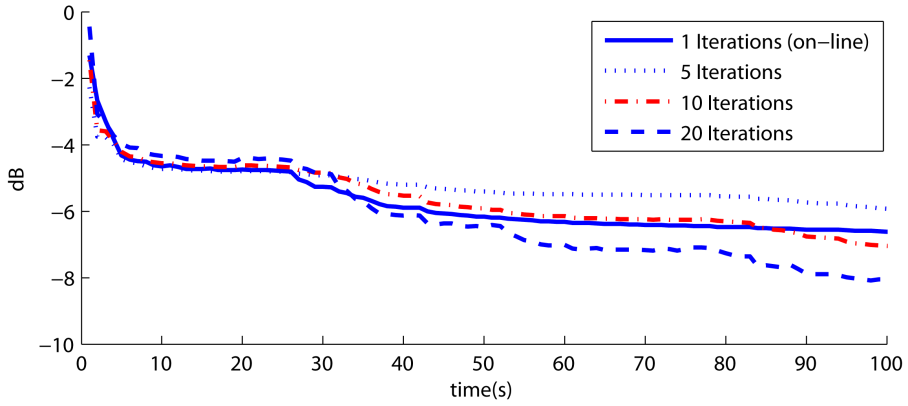
Figure 50: Performance of constrained and unconstrained SBSS (with 5 iterations).

Figure 51 shows the performance of SBSS with various number of iterations per batch. As the number of iterations is decreased, the adaptation process becomes more stable across time due to the reduction of the effect of the statistical bias, and the tERLE converge asymptotically to a higher value. However, the convergence rate is also decreased, which is expected since there is usually a trade-off between the convergence rate and the steady-state performance.

Figure 52 shows the effect of a near-end source signal on the SBSS performance. The signals were scaled appropriately to control the SNR between the loudspeakers and the



(a) True ERLE.



(b) Misalignment.

Figure 51: Performance of SBSS with different number of iterations (with de-mixing matrix constraint).

near-end signals. The EBR is defined as

$$\text{EBR} \equiv 10\log\frac{\langle r_i^2(t)\rangle_t}{\langle s_j^2(t)\rangle_t}, \quad \forall i,j \tag{156}$$

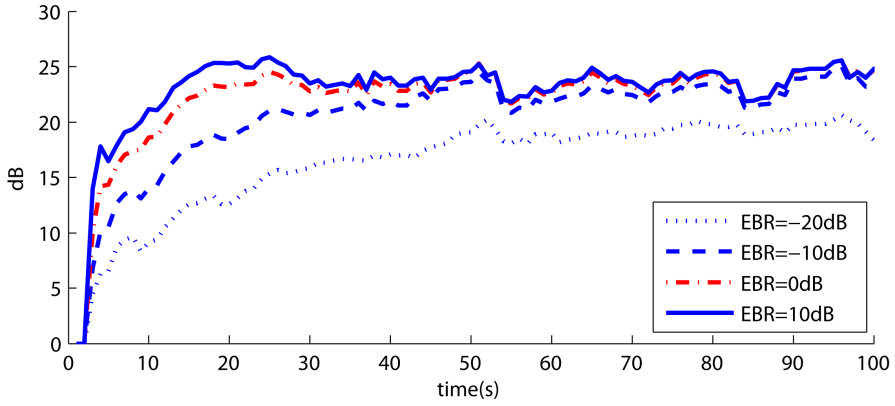
where r_i and s_j are the i^{th} and j^{th} reference and near-end signals, respectively, and $\langle\cdot\rangle_t$ indicates the averaging operator over time t. We can observe that even when the EBR is at -20 dB, which means that the acoustic echoes are almost inaudible with respect to the near-end signal, the SBSS algorithm is still able to achieve a high tERLE of about 20 dB. This experiment clearly demonstrates the robustness of the proposed SBSS algorithm to very noisy near-end mixing conditions and shows that the self-normalization introduced by the scaled natural gradient algorithm allows a stable adaptation that is mostly insensitive to differences in the input signal magnitude. This attractive characteristic was already observed in [27] for BSS and seems to be particularly effective also in the SBSS context.

Figure 53 shows the SBSS performance when the AWGN is added to the near-end microphone signals. We note that the performance does not vary considerably even with a very low SNR. It is worth noting that SBSS is capable of intrinsically adjusting to the presence of an interfering noise, which can be either a coherent source or an uncorrelated noise. In particular, all of the acoustic sources besides the acoustic echo at the near end can be considered as a single interfering source with given probability distribution. Thus the effectiveness of SBSS should depend also on the correct choice of the ICA contrast function $\Phi(x)$.

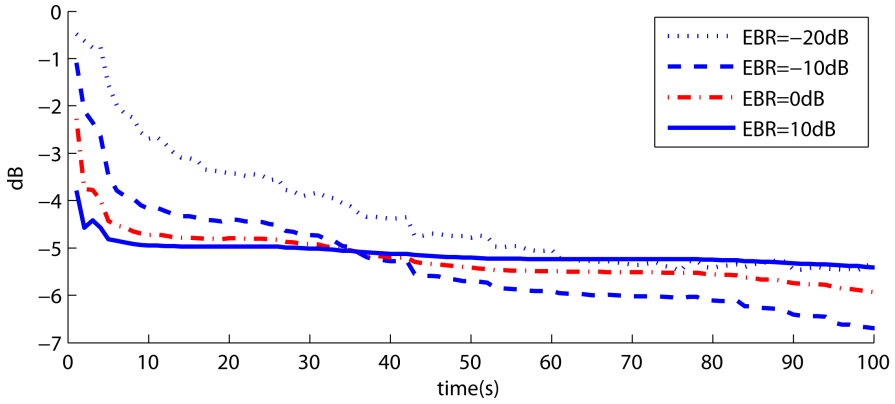
Figure 54 shows the SBSS performance with different FFT frame sizes. As expected, the performance decreases with the size of the FFT frame. In fact, the mixing system can be approximated by a linear instantaneous model at each frequency as long as the STFT windows size is sufficiently larger than the reverberation time.

Figure 55 shows the SBSS performance for different values of the ICA step-size η. Similarly to the influence of the number of iterations, the converge rate increases with the step-size at the cost of the steady-state performance.

Figure 56 shows the SBSS performance from using different number of iterations and STFT frame sizes for each batch. We gradually move from the full online implementation

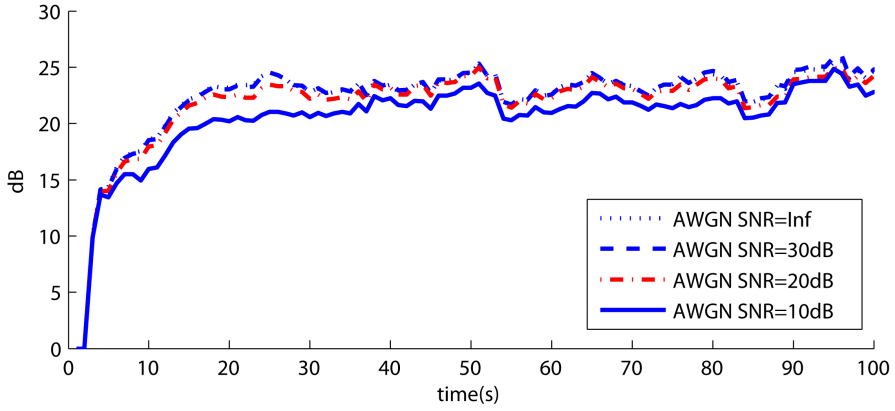


(a) True ERLE.

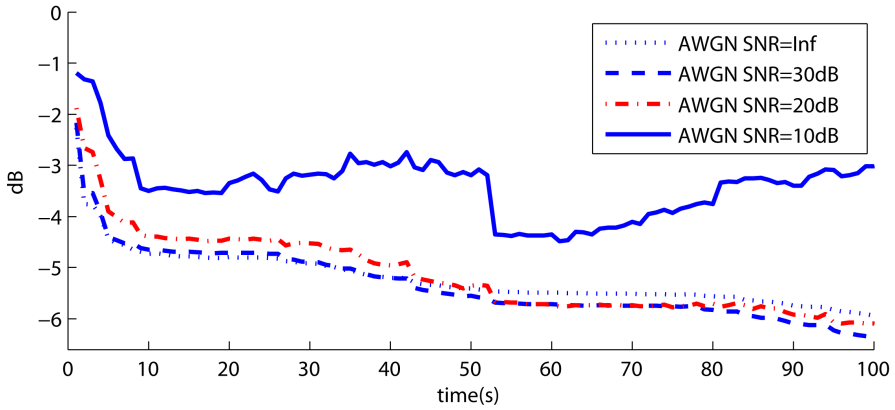


(b) Misalignment.

Figure 52: Performance of SBSS with different EBR (with de-mixing matrix constraint and 5 iterations).

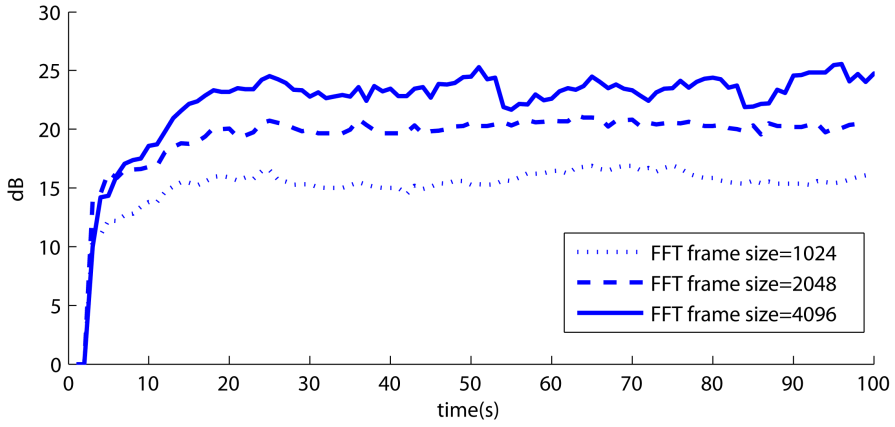


(a) True ERLE.

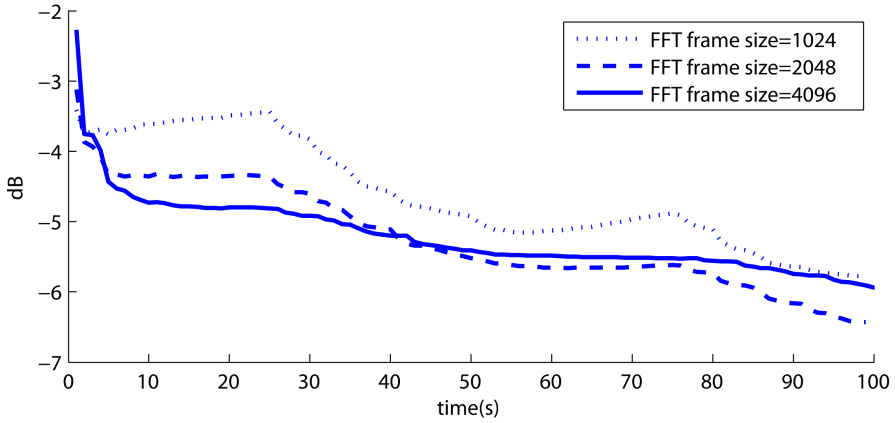


(b) Misalignment.

Figure 53: Performance of SBSS with AWGN noise at near-end microphones (with de-mixing matrix constraint and 5 iterations).

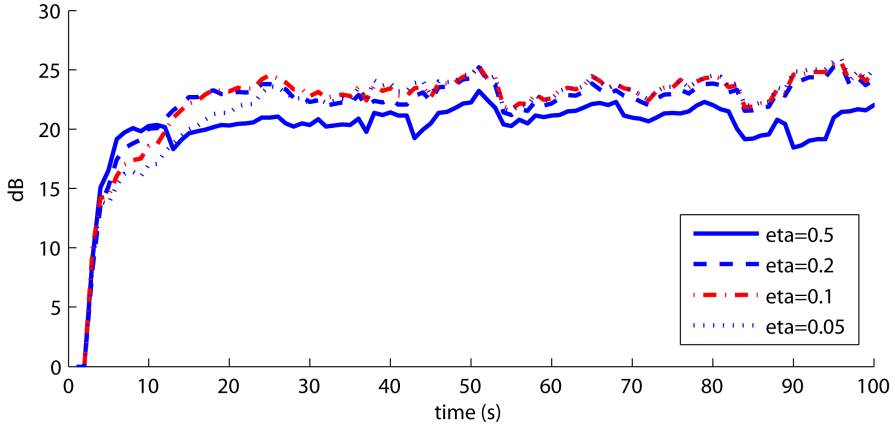


(a) True ERLE.

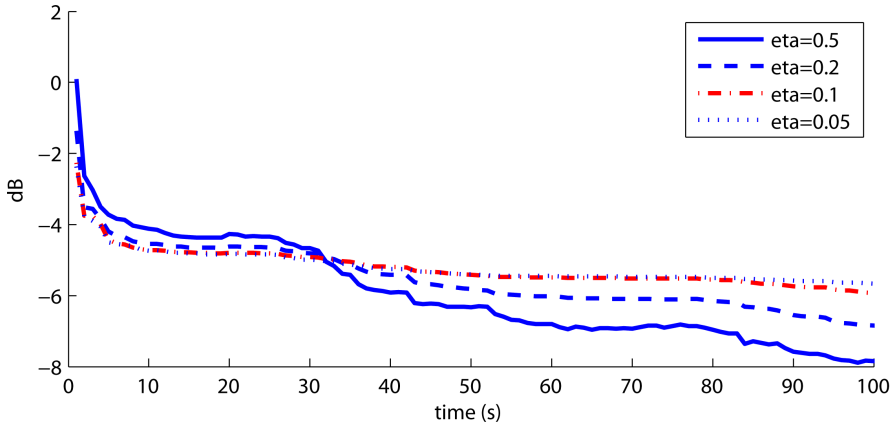


(b) Misalignment.

Figure 54: Performance of SBSS with different FFT frame size (with de-mixing matrix constraint and 5 iterations).



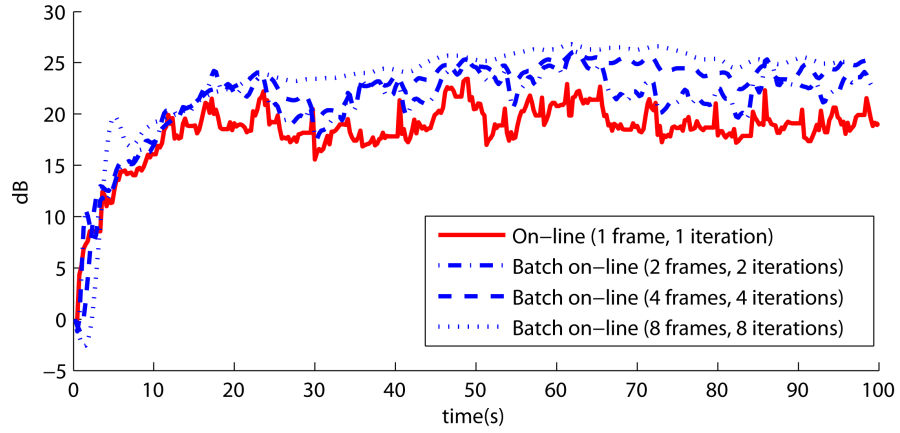
(a) True ERLE.



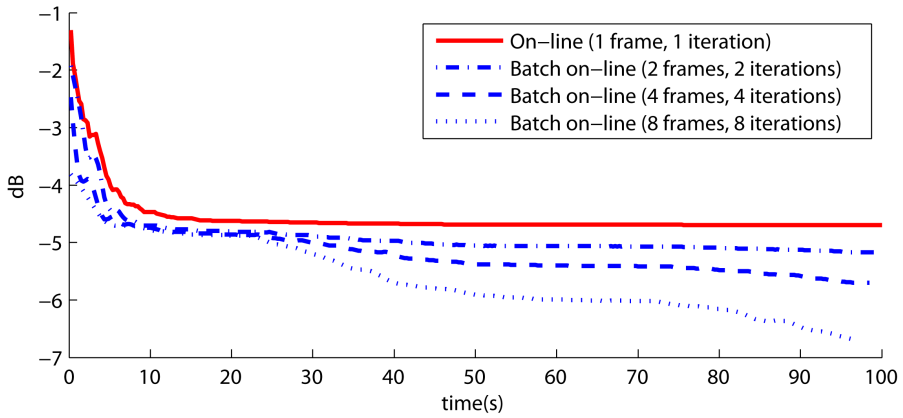
(b) Misalignment.

Figure 55: Performance of SBSS with different ICA step-size η (with de-mixing matrix constraint and 5 iterations).

(*i.e.*, covariance matrix $\mathbf{\Phi}$ is computed per frame and adapted by using (144) per iteration) to a batch-online implementation (*i.e.*, covariance matrix is averaged and adapted over larger set of multiple frames and iterations).



(a) True ERLE.



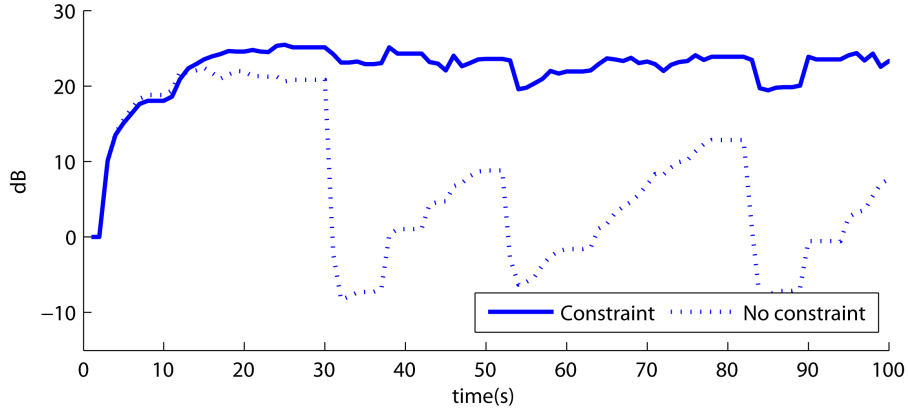
(b) Misalignment.

Figure 56: Performance of SBSS with online and batch-online implementations (with de-mixing matrix constraint and 5 iterations).

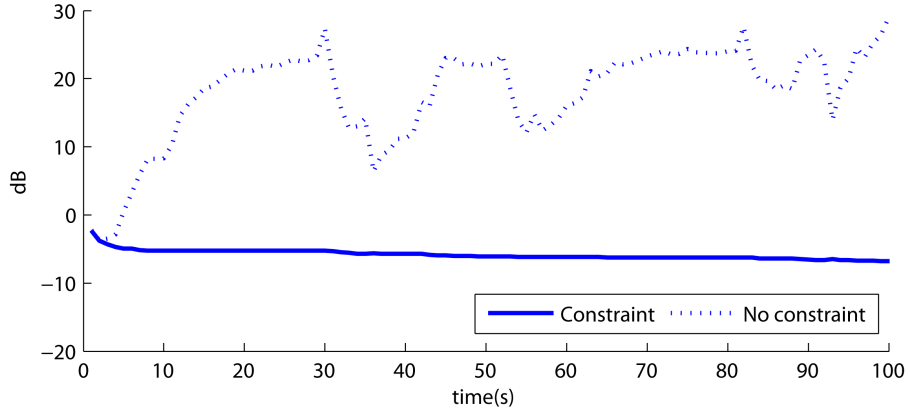
In both of the cases, the SBSS algorithm converges very quickly to a high tERLE. Similarly to the batch-online frequency-domain BSS in [85], it can be noted that the performance improves as we move from an online to a batch-online strategy. However, using batches with overly large number of frames may reduce the adaptability of the system to a variation in the mixing system, which again implies a trade-off between the convergence rate and the steady-state performance.

As we have previously discussed in Section 5.1.5 about the effect of the separation

matrix constraint on stability during changes in the source number or the mixing condition, we simulated another set of experiments with more realistic scenario where the number of both the near-end and far-end sources changes over time. Figure 57 shows the SBSS performance with and without the $\mathbf{W}_{22} = \mathbf{I}_R$ constraint when maximum of three sources were simultaneously active at the near end and the far end (*i.e.*, $0 \leq P \leq 2$ and $0 \leq Q \leq 3$), the number of the near-end and the far-end microphones were set to two, and AWGN with 80 dB SNR was added to the near-end microphone signals. The figure indicates that, similarly to the case discussed in Figure 50, the adaptation is stabilized with the matrix constraint regardless of the variation in the near-end and the far-end source numbers.



(a) True ERLE.



(b) Misalignment.

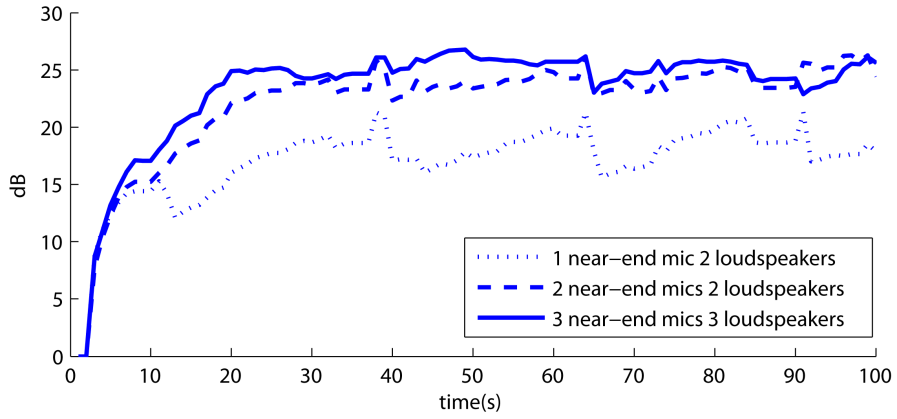
Figure 57: Performance of SBSS with variation in the active source number (with de-mixing matrix constraint and 5 iterations).

The effectiveness of the proposed framework can be better shown by evaluating the

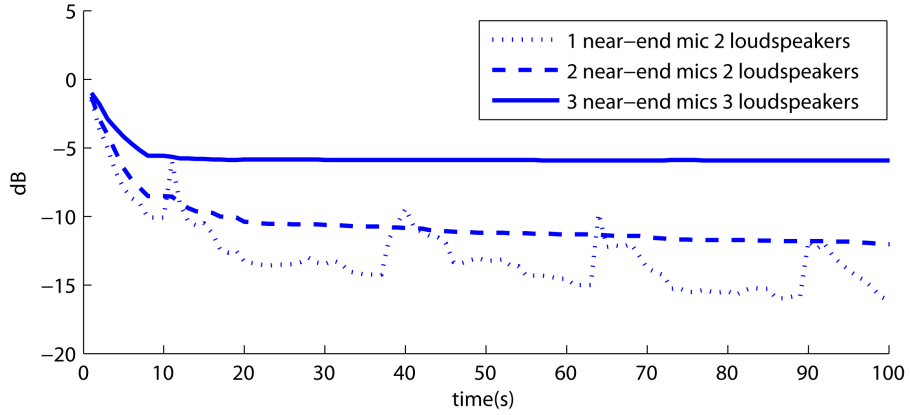
SBSS performance for several configurations of near-end and far-end microphones: (1) one near-end microphone and two far-end microphones; (2) two near-end microphones and two far-end microphones; and (3) three near-end microphones and three far-end microphones.

Figure 58 shows that the tERLE is higher for a larger number of near-end microphones while the misalignment is lower for a reduced number of the microphones. Such a result confirms that for SBSS, the echo cancellation performance does not correspond to the correct system identification. In fact, when many near-end microphones are used, the truly multi-channel nature of the SBSS adaptation allows the exploitation of spatial information through proper adaptation of the de-mixing sub-matrices $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$. This aspect becomes much more clear by observing Figure 59 that shows the SBSS performance when the near-end source separation is not applied, where $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$ are multiplied by $\mathbf{W}_{11}^{-1}$ and the AEC is performed by using causal mixing filters estimated by (152) and (153). As expected, the performance for the configurations with two and three near-end microphones become similar to the single microphone case. In fact, removing the effect of $\mathbf{W}_{11}$ is almost equivalent to separately applying the SBSS to each near-end microphone channel.

Finally, we compared the SBSS algorithm against a conventional multichannel AEC algorithm based on FBLMS [23], where we used a regularization procedure proposed in [52] to make FBLMS robust to the ambient background noise. The combined algorithm was implemented along with an ideal double-talk detector that froze adaptation by using the prior knowledge of near-end source activity. Real-world data were recorded in a room with a reverberation time of approximatively $T_{60} = 200$ ms, and the algorithms were tested with two near-end sources and two near-end loudspeakers (*i.e.*, two far-end microphones). The far-end and the near-end source numbers were varied over time, $0 \leq P, Q \leq 2$, where the near-end sources became active after 25 seconds. Changes in the near-end mixing condition were simulated by moving the microphones after 50 seconds and 80 seconds. Two adaptation strategies were used: (1) a batch-online adaptation with overlapped blocks of 1.024 second shifted by 0.128 second, and (2) an online STFT frame-by-frame adaptation. In both of the adaptation strategies, the signals were transformed through STFT and non-overlapping Hanning windows of 4096 samples, the covariance matrix was averaged over four frames

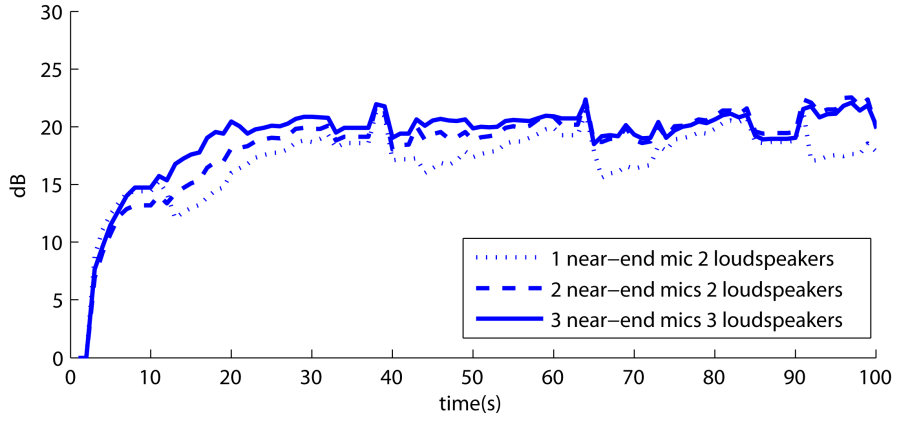


(a) True ERLE.

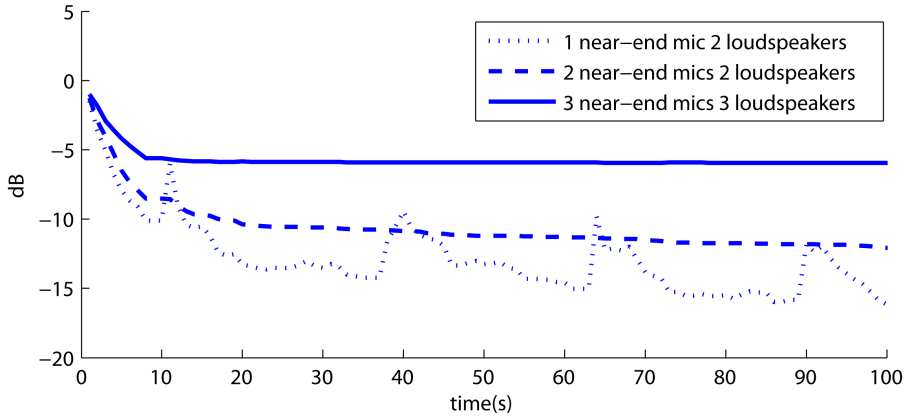


(b) Misalignment.

Figure 58: Performance of SBSS for different microphone configurations (with de-mixing matrix constraint and 5 iterations).



(a) True ERLE.



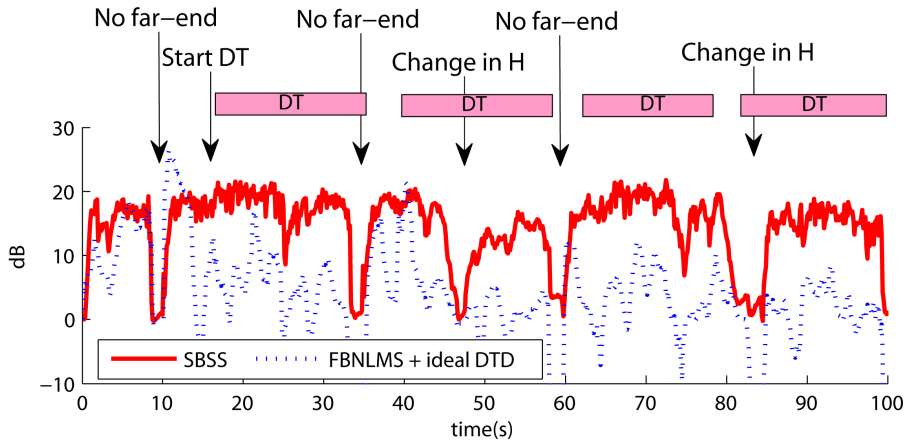
(b) Misalignment.

Figure 59: Performance of SBSS with different microphone configurations and without applying the near-end source separation (with de-mixing matrix constraint and 5 iterations).

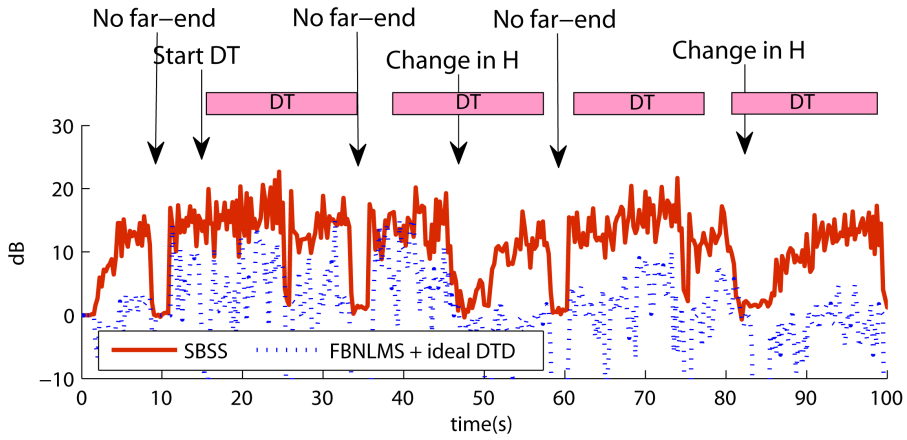
during the batch-online adaptation, and the solution for $\mathbf{W}$ was updated per iteration per block/frame.

Figure 60(a) shows the tERLE performance for the batch-online strategy where no smoothing was applied to the tERLE measurements. In the first 25 seconds, the FBLMS converges quickly to a high tERLE, but the echo cancellation performance is degraded when the near-end sources become active. On the other hand, while the SBSS exhibits slow convergence during the first 25 seconds, it clearly outperforms the FBLMS algorithm during double talk, most likely due to the scaled natural gradient algorithm. In particular, when the near-end source is active and when the mixing condition changes over time, the FBLMS algorithm is unable to track the variation in the room response while the SBSS algorithm has no such problem.

Figure 60(b) shows the results from the online adaptation strategy. In this case, the FBLMS algorithm displays consistently low convergence rate due to the noisy block-wise adaptation even when the near-end talkers are silent. The SBSS algorithm also suffers from a slower convergence rate than the corresponding batch-online implementation, but the adaptation is stable enough due to the scaled normalization of the natural gradient algorithm to ultimately provide acceptable echo cancellation performance. We note that the generalized covariance matrix used in the SBSS adaptation exploits the higher-order source statistics. Similarly to the case of BSS, any measures of higher-order source dependencies would be highly biased if the amount of data is limited. Therefore, in order to maximize the benefit of the SBSS structure and achieve the best overall echo cancellation performance, the batch-online adaptation is preferred over the online adaptation.



(a) True ERLE from batch-online implementation.



(b) True ERLE from on-line implementation.

Figure 60: Comparison between SBSS and FBLMS for a real-world scenario.

CHAPTER VI

CONCLUSIONS

We conclude by reviewing the main ideas presented in this dissertation.

First, we showed that the use of an error recovery nonlinearity (ERN) to "enhance" the error signal proves simple and effective in dealing with the robustness issue of acoustic echo cancellation (AEC) in the real world. The residual echo enhancement (REE), or error enhancement, paradigm arises from a very simple notion that reducing the effect of distortion remaining in the residual echo after the AEC and prior to the filter coefficients adaptation should provide for improved linear adaptive filtering performance in a noisy condition. The idea stems from the "system" approach to signal enhancement, where the system components should interact with one another mutually for the entire system's benefit. The ERN can be derived from well-established signal enhancement techniques based on the Bayesian statistical analysis. The combined technique evidently has close ties to the traditional noise-robust AEC schemes, namely the adaptive step-size and regularization procedures, and it can be readily utilized not only in the presence of an additive local noise but also when there is a nonlinear distortion on the acoustic echo due to, for example, a speech codec. Moreover, the ERN technique can be viewed as a generalization of the adaptive step-size procedure for non-Gaussian signals encountered in most real-world situations.

Most importantly, our study indicates that it is possible to advantageously circumvent the conventional practice of interrupting the filter adaptation in the presence of significant near-end interferences, *e.g.*, double talk. There is no need to freeze the filter adaptation entirely during the double-talk situation when the error enhancement procedure using a compressive ERN and a regularization procedure are combined together appropriately. Such a systematic paring allows the filter adaptation to be carried out continuously and recursively on a batch of very noisy data during the frequency-domain AEC. The "block-iterative" adaptation (BIA) in turn mitigates for the retarded convergence speed due to

an aggressive noise-robustness control enforced by the ERN. Recursive, batch-wise adaptation is nothing new for the blind source separation (BSS) algorithms based on independent component analysis (ICA) that are capable of unsupervised adaptation in the presence of multiple interfering signals. Indeed, the natural gradient algorithm optimized by ICA leads directly to the least mean square (LMS) algorithm and the ERN, where the combination is equivalent to semi-blind source separation (SBSS), *i.e.*, BSS when some source signals are given *a priori*, without the source separation.

Next, we successfully applied the REE technique to multi-channel AEC (MCAEC) in very noisy acoustic conditions. Other traditional techniques, such as double-talk detection and exponential weighting, were integrated into the AEC system to further improve the noise robustness and the overall cancellation performance. We proposed the decorrelation-by-resampling (DBR) technique with minimal signal distortion that effectively alleviates the non-uniqueness problem, which occurs due to inter-channel correlation when the LMS algorithm is used as a single-channel solution to MCAEC. We also developed the frequency-domain resampling (FDR) technique that is computationally efficient, as it relies on the Fast Fourier Transform for most of the interpolation work, and can be extended readily to more than two channels and higher sampling rates. The coherence measurement supports the intended design of DBR to smoothly decrease the coherence from low to high frequencies in order to minimize the signal distortion of the low frequency signal components, the process of which may degrade not only the audio quality but also the AEC performance itself. Simulation results indicate that the time-varying decorrelation via resampling is well suited for a frequency-domain MCAEC system with the REE procedure, as BIA used in conjunction with REE enables the recovery of lost cancellation performance due to the non-uniqueness problem and consequently reduces the necessity for aggressive decorrelation at low frequencies where most of the speech energy and information resides. In addition, we observed that BIA permits a natural recovery of the cancellation performance lost due to inter-block correlation, or inherent short-time correlation, of a speech signal when the multi-delay filter (MDF) is used.

As an extended realization of the system approach to decorrelation for AEC, the integration of the sub-band resampling (SBR) technique, obtained directly from FDR, with REE-based MCAEC leads to beneficial interaction between the decorrelation procedure and the AEC system as we showed through the sub-band analysis of the tERLE and misalignment. SBR allows selective decorrelation, measured in terms of the coherence, per frequency sub-band in order to, for example, leave the low frequency band unmodified to maintain high cancellation performance in that region naturally through BIA without the need for aggressive decorrelation. Such a selective decorrelation results in higher tERLE and lower misalignment overall than without decorrelation or with any other decorrelation procedures tested while achieving superior audio quality. SBR thus provides the flexibility to leave certain bands untouched and only resample other bands according to the desired AEC performance and the sound quality requirement.

Finally, we discussed many issues related to the implementation of an SBSS system that generalizes the MCAEC and is a truly multi-channel approach to AEC. We analyzed in detail the structure of the SBSS optimization and algorithmic issues in order to define a guideline for the implementation of robust SBSS systems. In particular, we proposed a constrained batch-online algorithm that stabilizes the adaptation process and improves the overall echo cancellation performance. Promising results show that in the simulated worst scenario case of a single far-end talker along with the non-uniqueness condition on the far-end mixing system, a stable adaptation is possible without the need of any decorrelation procedure on the reference signals. Furthermore, the adaptation is insensitive to the double-talk situation even when there are multiple near-end sources and acoustic echoes.

The problem of real-world AEC is best dealt with as a system issue, unlike the conventional single algorithm approaches in which the focus used to be unduly steered away from the robustness challenge. While we were first motivated by the intuitive notion of REE in the presence of noise and distortion, the technique is nonetheless a culmination of many diverse signal enhancement techniques. All of our findings thus far indicate strongly that the essential component in adaptive algorithms based on second-order statistics, namely the principle of orthogonality, can be elevated to the system level of an AEC problem. In a

rather interesting manner, such a connection allows us to relate AEC to source separation that seeks to maximize the independence, and thus implicitly the orthogonality, not only between the error signal and the far-end signal, but rather, among all signals involved. In other words, it promotes the conventional single-channel approach to signal enhancement with limited practical applicability to the global, system-level foundation where multiple information are available for the refinement of involved signals. This gives rise to an entirely new approach to the conventional AEC problem in a noisy and disruptive environment, where a single, idealized algorithm solution alone will not always be able to properly carry out its intended function as the acoustic mixing system becomes even more complex and prevalent in the future audio applications than ever before.

6.1 Contributions

We list below our specific contributions from this dissertation.

- Proposed the system approach to signal enhancement for tackling the real-world enhancement problems.
- Proposed the REE technique based on ICA for robust AEC performance in the presence of linear or nonlinear distortion at the near end.
- Proposed the system combination of REE, adaptive regularization procedure, and BIA in the frequency domain for noise-robust AEC without double-talk detection.
- Proposed applying BIA along with REE for a natural recovery of lost AEC performance due to inter-channel correlation during MCAEC and inter-block correlation during MDF.
- Proposed the DBR technique to directly alleviate the non-uniqueness problem with minimal distortion to audio quality and signal statistics.
- Proposed the FDR technique for expanded applicability of DBR, *e.g.*, reduction in the computational load and selective decorrelation via SBR.
- Proposed the SBSS as generalized, robust, and truly multi-channel solution to MCAEC.

Refer to the references chapter under the authors F. Nesta, E. Robledo-Arununcio, T. S. Wada, and J. Wung for two journal and fourteen conference publications that have come out so far from this work.

6.2 Future Research Suggestions

We suggest below other potential research subjects for future work.

- Improvement in the regularization and the REE procedures, *e.g.*, use of information from residual echo suppression (RES) to assist the error enhancement [143].
- Improvement in the DBR procedure, *e.g.*, perceptually hide the audio image fluctuation of high frequency components.
- Application of the proposed techniques to other adaptive filters, *e.g.*, APA and RLS.
- Application of the proposed techniques to the sampling rate mismatch problem, *e.g.*, use of FDR for sampling rate correction.
- Application of the proposed techniques to BSS and SBSS, *e.g.*, use of DBR for further improvement in source separation and MCAEC.
- Application of the system approach to other AEC structures, *e.g.*, sub-band AEC, nonlinear AEC.
- Application of the system approach to other aspects of AEC, *e.g.*, double-talk detection, convergence detection, RES [142].
- Application of the system approach to other signal enhancement problems, *e.g.*, feedback cancellation, active noise cancellation, generalized sidelobe canceller.

APPENDIX A

BAYESIAN ESTIMATION

A.1 MMSE Estimation

If $e = \bar{e} + v$, the convolution theorem [95] gives

$$p_e(e) = \int_{-\infty}^{\infty} p_v(e - \bar{e}) p_{\bar{e}}(\bar{e}) d\bar{e} \tag{157}$$

along with the property

$$p'_e(e) = \int_{-\infty}^{\infty} p_v(e - \bar{e}) p'_{\bar{e}}(\bar{e}) d\bar{e} = \int_{-\infty}^{\infty} p'_v(e - \bar{e}) p_{\bar{e}}(\bar{e}) d\bar{e}. \tag{158}$$

Substituting (48) into (49) gives

$$\begin{aligned} f_{\text{MMSE}}(e) &= \frac{\int_{-\infty}^{\infty} \bar{e}\, p_{e|\bar{e}}(e|\bar{e}) p_{\bar{e}}(\bar{e}) d\bar{e}}{\int_{-\infty}^{\infty} p_{e|\bar{e}}(e|\bar{e}) p_{\bar{e}}(\bar{e}) d\bar{e}} \\ &= \frac{\int_{-\infty}^{\infty} \bar{e}\, p_v(e - \bar{e}) p_{\bar{e}}(\bar{e}) d\bar{e}}{\int_{-\infty}^{\infty} p_v(e - \bar{e}) p_{\bar{e}}(\bar{e}) d\bar{e}} = e - \frac{\int_{-\infty}^{\infty} v\, p_{\bar{e}}(e - v) p_v(v) dv}{\int_{-\infty}^{\infty} p_{\bar{e}}(e - v) p_v(v) dv}. \end{aligned} \tag{159}$$

Then applying (52) and (54) to (159) gives (59) for $\bar{e} \sim \mathit{Gaussian}(0, \sigma_{\bar{e}})$ and any v, (63) for $v \sim \mathit{Gaussian}(0, \sigma_v)$ and any $\bar{e}$, and (67) for $\bar{e} \sim \mathit{Gaussian}(0, \sigma_{\bar{e}})$ and $v \sim \mathit{Gaussian}(0, \sigma_v)$. Furthermore, applying (58) to (59) gives (61) for $v \sim \mathit{Laplacian}(0, \alpha_v)$, and applying (58) to (63) gives (65) for $\bar{e} \sim \mathit{Laplacian}(0, \alpha_{\bar{e}})$. Similar derivations are provided in [42, 75].

A.2 MAP Estimation

Let $f_s(s) = -\log p_s(s)$. Then (50) can be rewritten as

$$\begin{aligned} f_{\text{MAP}}(e) &= \underset{\bar{e}}{\operatorname{argmin}} \{f_v(e - \bar{e}) + f_{\bar{e}}(\bar{e})\} \\ &= \{\hat{\bar{e}} : [f_v(e - \bar{e}) + f_{\bar{e}}(\bar{e})]'_{\bar{e}=\hat{\bar{e}}} = 0\} \\ &= \{\hat{\bar{e}} : -\phi_v(e - \hat{\bar{e}}) + \phi_{\bar{e}}(\hat{\bar{e}}) = 0\}. \end{aligned} \tag{160}$$

The MAP nonlinearity is obtained by solving $\phi_{\bar{e}}(\hat{\bar{e}}) = \phi_v(e - \hat{\bar{e}})$ for $\hat{\bar{e}}$ as a function of e. First, if $\bar{e} \sim \mathit{Gaussian}(0, \sigma_{\bar{e}})$, then (60) is obtained for any v by applying (53) to (160) to

get

$$\hat{\bar{e}}/\sigma_{\bar{e}}^2 = \phi_v(e - \hat{\bar{e}}). \tag{161}$$

Applying (56) to (161) gives (62) for $v \sim Laplacian(0, \alpha_v)$. Next, if $v \sim Gaussian(0, \sigma_v)$, then (64) is obtained for any $\bar{e}$ by applying (53) to (160) to get

$$\phi_{\bar{e}}(\hat{\bar{e}}) = (e - \hat{\bar{e}})/\sigma_v^2. \tag{162}$$

Applying (56) to (162) gives (66) for $\bar{e} \sim Laplacian(0, \alpha_{\bar{e}})$. Finally, if $\bar{e} \sim Gaussian(0, \sigma_{\bar{e}})$ and $v \sim Gaussian(0, \sigma_v)$, then applying (53) to (160) gives (67). Similar derivations are provided in [55, 42].

APPENDIX B

FISHER INFORMATION

Non-Gaussianity for some PDF p_s can be measured through the Fisher information [55]

$$I(s) = E\{[\phi_s(s)]^2\} \tag{163}$$

with the property

$$I(s/a) = a^2 I(s) \tag{164}$$

for a constant parameter a. Given $e = \bar{e} + v$, it can be shown by following the proof in [55] that the relative improvement in noise reduction by using (60) over the linear Wiener filtering rule of (67) when $\bar{e} \sim Gaussian(0, \sigma_{\bar{e}})$ but v is zero-mean non-Gaussian distributed with the variance γ_v^2 is

$$R_{effective}^{\mathrm{GA}} = [I(v/\sigma_{\bar{e}}) + I(v/\gamma_v) - 1]\sigma_{\bar{e}}^2/\gamma_v^2 + o(\sigma_{\bar{e}}^2) \tag{165}$$

for $\sigma_{\bar{e}}^2 < \gamma_v^2$, where "$G$" stands for Gaussian and "A" for any distribution. Similarly, the relative improvement by using (64) over (71) when $v \sim Gaussian(0, \sigma_v)$ but $\bar{e}$ is zero-mean non-Gaussian distributed with the variance $\gamma_{\bar{e}}^2$ is

$$R_{effective}^{\mathrm{AG}} = [I(\bar{e}/\gamma_{\bar{e}}) + I(\bar{e}/\sigma_v) - 1]\sigma_v^2/\gamma_{\bar{e}}^2 + o(\sigma_v^2) \tag{166}$$

for $\sigma_v^2 < \gamma_{\bar{e}}^2$. In a set of PDF with unit variance, (163) is minimized and is equal to 1 for the Gaussian PDF. Therefore, (165) and (166) indicate improved noise suppression when either one of $\bar{e}$ and v is Gaussian distributed while the other is non-Gaussian distributed.

APPENDIX C

GENERALIZED DECORRELATION

Assuming that $\mathbf{x} = \{x_j\}$ is a vector of zero-mean random variables of length N and that $\mathbf{A} = \{a_{ij}\}$ is an $N \times N$ matrix with constant elements, the statistical moment of order $\mathtt{u}$ for the i^{th} element of $\mathbf{Ax}$ is given by

$$E\left\{(\sum_{j=1}^{N} a_{ij}x_j)^{\mathtt{u}}\right\} \quad \forall i. \tag{167}$$

By using the multinomial expansion, (167) can be rewritten as

$$E\left\{\sum_{\substack{l_1,l_2,\ldots,l_N \geq 0 \\ l_1+l_2+\ldots+l_N=\mathtt{u}}} \frac{\mathtt{u}!}{l_1!..l_N!} \prod_{j=1}^{N} (a_{ij}x_j)^{l_j}\right\} \quad \forall i. \tag{168}$$

If the elements in $\mathbf{x}$ are mutually independent, (168) reduces to

$$E\left\{\sum_{j}(a_{ij}x_j)^{\mathtt{u}}\right\} = \sum_{j} a_{ij}^{\mathtt{u}} E\{x_j^{\mathtt{u}}\} \quad \forall i. \tag{169}$$

We can then generalize that

$$E\{(\mathbf{Ax})^{\mathtt{u}}\} = \mathbf{A}^{\mathtt{u}} E\{\mathbf{x}^{\mathtt{u}}\}, \tag{170}$$

where $\mathbf{x}^{\mathtt{u}}$ and $\mathbf{A}^{\mathtt{u}}$ indicate the raising of each element of the vector $\mathbf{x}$ and of the matrix $\mathbf{A}$ to the power $\mathtt{u}$. By the property of the covariance of linear combinations of variables, we know that if the random variables in $\mathbf{x}$ are independent, then given the $N \times N$ matrices $\mathbf{A}$ and $\mathbf{B}$, we have

$$E\{\mathbf{Axx}^H\mathbf{B}\} = \mathbf{A}E\{\mathbf{xx}^H\}\mathbf{B}. \tag{171}$$

By using (168) and following the derivation of (170), it is possible to generalize (171) for higher-order moments as

$$E\{(\mathbf{Ax})^{\mathtt{u}}\mathbf{x}^H\mathbf{B}\} = \mathbf{A}^{\mathtt{u}} E\{\mathbf{x}^{\mathtt{u}}\mathbf{x}^H\}\mathbf{B}. \tag{172}$$

REFERENCES

[1] Aboulnasr, T. and Mayyas, K., "A robust variable step-size LMS-type algorithm: Analysis and simulations," *IEEE Trans. Signal Process.*, vol. 45, pp. 631–639, Mar. 1997.

[2] Al-Naffouri, T. Y. and Sayed, A. H., "Adaptive filters with error nonlinearities: Mean-square analysis and optimum design," *EURASIP Applied Signal Process.*, vol. 2001, pp. 192–205, Oct. 2001.

[3] Amari, S., "Natural gradient works efficiently in learning," *Neural Computat.*, vol. 10, no. 2, pp. 251–276, 1998.

[4] Amari, S.-I., Chen, T.-P., and Cichocki, A., "Nonholonomic orthogonal learning algorithms for blind source separation," *Neural Computat.*, vol. 12, no. 6, pp. 1463–1484, 2000.

[5] Ang, W.-P. and Farhang-Boroujeny, B., "A new class of gradient adaptive step-size LMS algorithms," *IEEE Trans. Signal Process.*, vol. 49, pp. 805–810, Apr. 2001.

[6] Bell, A. J. and Sejnowski, T. J., "An information-maximization approach to blind separation and blind deconvolution," *Neural Computat.*, vol. 7, pp. 1129–1159, Nov. 1995.

[7] Benesty, J. and Gänsler, T., "A robust fast recursive least squares adaptive algorithm," in *Proc. IEEE ICASSP*, vol. 6, pp. 3785–3788, May 2001.

[8] Benesty, J. and Gänsler, T., "On data-reuse adaptive algorithms," in *Proc. IWAENC*, Sep. 2003.

[9] Benesty, J., Gänsler, T., Morgan, D. R., Sondhi, M. M., and Gay, S. L., *Advances in Network and Acoustic Echo Cancellation*. Springer, 2001.

[10] Benesty, J., Morgan, D. R., and Cho, J. H., "A new class of doubletalk detectors based on cross-correlation," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 168–172, Mar. 2000.

[11] Benesty, J., Morgan, D. R., Hall, J. L., and Sondhi, M. M., "Stereophonic acoustic echo cancellation using nonlinear transformations and comb filtering," in *Proc. IEEE ICASSP*, vol. 6, pp. 3673–3676, May 1998.

[12] Benesty, J., Morgan, D. R., and Sondhi, M. M., "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 156–165, Mar. 1998.

[13] Benveniste, A., Metivier, M., and Priouret, P., *Adaptive Algorithms and Stochastic Approximation*. Springer, 1990.

[14] Bershad, N. J., "On error-saturation nonlinearities in LMS adaptation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, pp. 440–452, Apr. 1988.

[15] Birkett, A. N. and Goubran, R. A., "Acoustic echo cancellation using NLMS-neural network structures," in *Proc. IEEE ICASSP*, vol. 5, pp. 3035–3038, May 1995.

[16] Birkett, A. N. and Goubran, R. A., "Limitations of handsfree acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects," in *Proc. IEEE WASPAA*, pp. 103–106, Oct. 1995.

[17] Borrallo, J. M. P. and Otero, M. G., "On the implementation of a partitioned block frequency domain adaptive filter (pbfdaf) for long acoustic echo cancellation," *Signal Process.*, vol. 27, pp. 301–315, Jun. 1992.

[18] Buchner, H., Benesty, J., Gänsler, T., and Kellermann, W., "An outlier-robust extended multidelay filter with application to acoustic echo cancellation," in *Proc. IWAENC*, pp. 19–22, Sep. 2003.

[19] Buchner, H., Benesty, J., and Kellermann, W., "Generalized multichannel frequency-domain adaptive filtering: Efficient realization and application to hands-free speech communication," *Signal Process.*, vol. 85, pp. 549–570, Mar. 2005.

[20] Carter, G., "Coherence and time delay estimation," *Proc. IEEE*, vol. 75, pp. 236–255, Feb. 1987.

[21] Cichocki, A. and Amari, S., *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. John Wiley & Sons, 2002.

[22] Cichocki, A., Sabala, I., and Amari, S., "Intelligent neural networks for blind signal separation with unknown number of sources," in *Proc. CEIS*, pp. 148–154, 1998.

[23] Clark, G. A., Parker, S. R., and Mitra, S. K., "A unified approach to time- and frequency-domain realization of FIR adaptive digital filters," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-31, pp. 1073–1083, Oct. 1983.

[24] Comon, P., "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, pp. 287–314, Apr. 1994.

[25] Diethorn, E. J., "Improved decision logic for two-path echo cancelers," in *Proc. IWAENC*, Sep. 2001.

[26] Douglas, S. C., "Generalized gradient adaptive step sizes for stochastic gradient adaptive filters," in *Proc. IEEE ICASSP*, vol. 2(9), pp. 1396–1399, May 1995.

[27] Douglas, S. C. and Gupta, M., "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation," in *Proc. IEEE ICASSP*, vol. 2(5), pp. 637–640, Apr. 2007.

[28] Douglas, S. C. and Meng, T. H.-Y., "Stochastic gradient adaptation under general error criteria," *IEEE Trans. Signal Process.*, vol. 42, pp. 1335–1351, June 1994.

[29] Duttweiler, D. L., "A twelve-channel digital echo canceller," *IEEE Trans. Commun.*, vol. COM-26, pp. 647–653, May 1978.

[30] Duttweiler, D. L., "Adaptive filter performance with nonlinearities in the correlation multiplier," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-30, pp. 578–586, Aug. 1982.

[31] Faller, C., Gänsler, T., and Vetterli, M., "Two stage estimation for the echo paths in stereophonic acoustic echo cancellation," in *Proc. IWAENC*, pp. 157–160, Sep. 2006, paper no. 40.

[32] Fozunbal, M., Kalker, T., and Schafer, R. W., "Multi-channel echo control by model learning," in *Proc. IWAENC*, Sep. 2008, paper no. 9014.

[33] Gänsler, T., "A double-talk resistant subband echo canceller," *Signal Process.*, vol. 65, pp. 89–101, Feb. 1998.

[34] Gänsler, T. and Benesty, J., "Stereophonic acoustic echo cancellation and two-channel adaptive filtering: an overview," in *J. Adapt. Control Signal Process.*, vol. 14, pp. 565–586, 2000.

[35] Gänsler, T. and Benesty, J., "A frequency-domain double-talk detector based on a normalized cross-correlation vector," *Signal Process.*, vol. 81, pp. 1783–1787, Aug. 2001.

[36] Gänsler, T. and Benesty, J., "New insights into the stereophonic acoustic echo cancellation problem and an adaptive nonlinearity solution," *IEEE Trans. Speech Audio Process.*, vol. 10, pp. 257–267, July 2002.

[37] Gänsler, T. and Benesty, J., "The fast normalized cross-correlation double-talk detector," *Signal Process.*, vol. 86, pp. 1124–1139, June 2006.

[38] Gänsler, T., Gay, S. L., Sondhi, M. M., and Benesty, J., "Double-talk robust fast converging algorithms for network echo cancellation," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 656–663, Nov. 2000.

[39] Gänsler, T., Hansson, M., Ivarson, C. J., and Salomonsson, G., "A double-talk detector based on coherence," *IEEE Trans. Commun.*, vol. 44, pp. 1421–1427, Nov. 1996.

[40] Gay, S. L. and Benesty, J., *Acoustic Signal Processing for Telecommunication*. Kluwer Academic, 2000.

[41] Gazor, S. and Zhang, W., "Speech probability distribution," *IEEE Signal Process. Letters*, vol. 10, pp. 204–207, Jul. 2003.

[42] Gazor, S. and Zhang, W., "Speech enhancement employing Laplacian-Gaussian mixture," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 896–904, Sep. 2005.

[43] Guérin, A., Faucon, G., and Bouquin-Jeannes, R. L., "Nonlinear acoustic echo cancellation based on Volterra filters," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 672–683, Nov. 2003.

[44] Gustaffson, T., Rao, B. D., and Trivedi, M., "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 791–803, Nov. 2003.

[45] Hänsler, E. and Schmidt, G. U., "Hands-free telephonesjoint control of echo cancellation and postfiltering," *Signal Process.*, vol. 80, pp. 2295–2305, Oct. 2000.

[46] Hänsler, E. and Schmidt, G. U., *Acoustic Echo and Noise Control: A Practical Approach*. John Wiley & Sons, 2004.

[47] Hassibi, B., Sayed, A. H., and Kailath, T., "H^∞ optimality of the LMS algorithm," *IEEE Trans. Signal Process.*, vol. 44, pp. 267–280, Feb. 1996.

[48] Haykin, S., *Blind Deconvolution*. Prentice Hall, 1994.

[49] Haykin, S., *Adaptive Filter Theory*. Prentice Hall, 4th ed., 2002.

[50] Heckmann, M., Vogel, J., and Kroschel, K., "Frequency selective step-size control for acoustic echo cancellation," in *Proc. EURASIP EUSIPCO*, pp. 1855–1858, Sep. 2000.

[51] Herre, J., Buchner, H., and Kellermann, W., "Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement," in *Proc. IEEE ICASSP*, vol. I, pp. 17–20, Apr. 2007.

[52] Hirano, A. and Sugiyama, A., "A noise-robust stochastic gradient algorithm with an adaptive step-size for mobile hands-free telephones," in *Proc. IEEE ICASSP*, vol. 2, pp. 1392–1395, May 1995.

[53] Huang, Y. and Benesty, J., eds., *Audio Signal Processing for Next-Generation Multimedia Communication Systems*. Kluwer Academic, 2004.

[54] Huber, P. J., *Robust Statistics*. John Wiley & Sons, 1981.

[55] Hyvärinen, A., "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neural Computat.*, vol. 11, pp. 1739–1768, Oct. 1999.

[56] Hyvärinen, A., Karhunen, J., and Oja, E., *Independent Component Analysis*. John Wiley & Sons, 2001.

[57] Ikeda, K. and Sakamoto, R., "Convergence analyses of stereo acoustic echo cancelers with preprocessing," *IEEE Trans. Signal Process.*, vol. 51, pp. 1324–1334, May 2003.

[58] Inst., E. T. S., "ETSI TS 126 104 V6.1.0: ANSI-C code for the floating-point Adaptive Multi-Rate (AMR) speech codec," 2004.

[59] Iqbal, M. A., Stokes, J., and Grant, S. L., "Normalized double-talk detection based on microphone and AEC error cross-correlation," in *Proc. IEEE ICME*, vol. 2, pp. 360–363, July 2007.

[60] Joho, M., Mathis, H., and Moschytz, G. S., "Combined blind/nonblind source separation based on the natural gradient," *IEEE Signal Process. Letters*, vol. 8, pp. 236–238, Aug. 2001.

[61] Kallinger, M., Mertins, A., and Kammeyer, K. D., "Enhanced double-talk detection based on pseudo-coherence in stereo," in *Proc. IWAENC*, pp. 177–180, Sep. 2005.

[62] Khong, A. W. H., Benesty, J., and Naylor, P. A., "Effect of interchannel coherence on conditioning and misalignment performance for stereo acoustic echo cancellation," in *Proc. IEEE ICASSP*, vol. 5, pp. 265–268, May 2006.

[63] Khong, A. W. H., Benesty, J., and Naylor, P. A., "Stereophonic acoustic echo cancellation: Analysis of the misalignment in the frequency domain," *IEEE Signal Process. Letters*, vol. 13, pp. 33–36, Jan. 2006.

[64] Khong, A. W. H. and Naylor, P. A., "Stereophonic acoustic echo cancellation employing selective-tap adaptive algorithms," *IEEE Trans. Audio Speech Language Process.*, vol. 14, pp. 785–796, May 2006.

[65] Koike, S., "A class of adaptive step-size control algorithms for adaptive filters," *IEEE Trans. Signal Process.*, vol. 50, pp. 1315–1326, June 2002.

[66] Kuech, F. and Kellermann, W., "Orthogonalized power filters for nonlinear acoustic echo cancellation," *Signal Process.*, vol. 86, pp. 1168–1181, June 2006.

[67] Kuech, F. and Kellermann, W., "Nonlinear residual echo suppression using a power filter model of the acoustic echo path," in *Proc. IEEE ICASSP*, vol. 1, pp. 73–76, Apr. 2007.

[68] Lee, T.-W., *Independent Component Analysis: Theory and Applications*. Kluwer Academic, 1998.

[69] Lehmann, E. and Johansson, A., "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. America*, vol. 124(1), pp. 269–277, July 2008.

[70] MacKay, D. J. C., *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[71] Mader, A., Puder, H., and Schmidt, G. U., "Step-size control for acoustic echo cancellation filters - an overview," *Signal Process.*, vol. 80, pp. 1697–1719, Sep. 2000.

[72] Makino, S., Kaneda, Y., and Koizumi, N., "Exponentially weighted step-size NLMS adaptive filter based on the statistics of a room impulse response," *IEEE Trans. Speech Audio Process.*, vol. 1, pp. 101–108, Jan. 1993.

[73] Makino, S., Lee, T.-W., and Sawada, H., eds., *Blind Speech Separation*. Springer, 2007.

[74] Mandic, D. P., "A generalized normalized gradient descent algorithm," *IEEE Signal Process. Letters*, vol. 11, pp. 115–118, Feb. 2004.

[75] Martin, R., "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 845–856, Sep. 2005.

[76] Mathews, V. J. and Cho, S. H., "Improved convergence analysis of stochastic gradient adaptive filters using the sign algorithm," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-35, pp. 450–454, Apr. 1987.

[77] Mathews, V. J. and Xie, Z., "A stochastic gradient adaptive filter with gradient adaptive stepsize," *IEEE Trans. Signal Process.*, vol. 41, pp. 2075–2087, June 1993.

[78] Matsuoka, K. and Nakashima, S., "Minimal distortion principle for blind source separation," in *Proc. ICA*, vol. 3, pp. 722–727, Dec. 2001.

[79] Miyabe, S., Takatani, T., Saruwatari, H., Shikano, K., and Tatekura, Y., "Barge-in and noise-free spoken dialogue interface based on sound field control and semi-blind source separation," in *Proc. EURASIP EUSIPCO*, pp. 232–236, Sep. 2007.

[80] Miyabe, S., *Barge-in Robust Spoken Dialogue Interface Using Multichannel Sound Field Control and Array Signal Processing*. PhD thesis, Nara Institute of Science and Technology, Nara, Japan, Sep. 2007.

[81] Moon, T. K., "The expectation-maximization algorithm," *IEEE Signal Process. Magazine*, vol. 13, pp. 47–60, Nov. 1996.

[82] Moulines, E., Ait-Amrane, O., and Grenier, Y., "The generalized multidelay adaptive filter: Structure and convergence analysis," *IEEE Trans. Signal Process.*, vol. 43, pp. 14–28, Jan. 1995.

[83] Moulines, E., Ait-Amrane, O., and Grenier, Y., "The generalized multidelay adaptive filter: Structure and convergence analysis," *IEEE Trans. Signal Process.*, vol. 43, pp. 14–28, Jan. 1995.

[84] Moulines, E., Cardoso, J.-F., and Gassiat, E., "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," in *Proc. IEEE ICASSP*, pp. 3617–3620, Apr. 1997.

[85] Mukai, R., Sawada, H., Araki, S., and Makino, S., "Real-time blind source separation an DOA estimation using small 3-D microphone array," in *Proc. IWAENC*, pp. 45–48, Sep. 2005.

[86] Murata, N. and Ikeda, S., "An on-line algorithm for blind source separation on speech signals," in *Proc. NOLTA*, vol. 3, pp. 923–926, Sep. 1998.

[87] Murata, N., Ikeda, S., and Ziehe, A., "An approach to blind source separation based on temporal structure of speech signals," *Neural Computat.*, vol. 41, pp. 1–24, Aug 2001.

[88] Nesta, F., Omologo, M., and Svaizer, P., "A novel robust solution to the permutation problem based on a joint multiple TDOA estimation," in *Proc. IWAENC*, Sep. 2008, paper no. 9023.

[89] Nesta, F., Wada, T. S., and Juang, B.-H., "Batch-online semi-blind source separation applied to multi-channel acoustic echo cancellation," *IEEE Trans. Audio Speech Language Process.*, vol. 19, pp. 583–599, Mar. 2011.

[90] Nesta, F., Wada, T. S., Miyabe, S., and Juang, B.-H., "On the non-uniqueness problem and the semi-blind source separation," in *Proc. IEEE WASPAA*, pp. 101–104, Oct. 2009.

[91] Nitsch, B. H., "A frequency-selective stepfactor control for an adaptive filter algorithm working in the frequency domain," *Signal Process.*, vol. 80, pp. 1733–1745, Sep. 2000.

[92] Ochiai, K., Araseki, T., and Ogihara, T., "Echo canceler with two echo path models," *IEEE Trans. Commun.*, vol. COM-25, pp. 589–595, June 1977.

[93] Oppenheim, A. V., Schafer, R. W., and Buck, J. R., *Discrete-Time Signal Processing*. Prentice Hall, 2nd ed., 1999.

[94] Ozeki, K. and Umeda, T., "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," *Trans. Inst. Elect. Info. Commun. Engineers*, vol. J67-A, pp. 126–132, Feb. 1984.

[95] Papoulis, A. and Pillai, S. U., *Probability, Random Variables and Stochastic Processes*. McGraw Hill, 4th ed., 2002.

[96] Park, H.-M., Oh, S.-H., and Lee, S.-Y., "On adaptive noise cancelling based on independent component analysis," *Electronics Letters*, vol. 38, pp. 832–833, Jul. 2002.

[97] Pazaitis, D. I. and Constantinides, A. G., "A novel kurtosis driven variable step-size adaptive algorithm," *IEEE Trans. Signal Process.*, vol. 47, pp. 864–872, Mar. 1999.

[98] Pham, D.-T., Servière, C., and Boumaraf, H., "Blind separation of speech mixtures based on nonstationarity," in *Proc. IEEE ISSPA*, vol. 2, pp. 73–76, July 2003.

[99] Poor, H. V., *An Introduction to Signal Detection and Estimation*. Springer, 2nd ed., 1994.

[100] Robledo-Arununcio, E., Wada, T. S., and Juang, B.-H., "On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation," in *Proc. IEEE WASPAA*, pp. 34–37, Oct. 2007.

[101] Saruwatari, H., Kurita, S., Takeda, K., Itakura, F., Nishikawa, T., and Shikano, K., "Blind source separation combining independent component analysis and beamforming," *EURASIP Applied Signal Process.*, no. 1, pp. 1135–1146, 2003.

[102] Sawada, H., Araki, S., Mukai, R., and Makino, S., "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. Audio Speech Language Process.*, vol. 15, pp. 1592–1604, July 2007.

[103] Schobben, D. W. E. and Sommen, P. W., "A frequency domain blind signal separation method based on decorrelation," *IEEE Trans. Signal Process.*, vol. 50, pp. 1855–1865, Aug. 2002.

[104] Sethares, W. A., "Adaptive algorithms with nonlinear data and error functions," *IEEE Trans. Signal Process.*, vol. 40, pp. 2199–2206, Sep. 1992.

[105] Shi, K., *Nonlinear Acoustic Echo Cancellation*. PhD thesis, Georgia Institute of Technology, Atlanta, GA, USA, Dec. 2008.

[106] Shi, K., Ma, X., and Zhou, G. T., "A double-talk detector based on generalized mutual information for stereophonic acoustic echo cancellation in the presence of nonlinearity," in *Proc. IEEE ACSSC*, Oct. 2008.

[107] Shi, K., Ma, X., and Zhou, G. T., "A mutual information based double-talk detector for nonlinear systems," in *Proc. IEEE CISS*, pp. 356–360, Mar. 2008.

[108] Shi, K., Ma, X., and Zhou, G. T., "A residual echo suppression technique for systems with nonlinear acoustic echo paths," in *Proc. IEEE ICASSP*, pp. 257–260, Apr. 2008.

[109] Shi, K., Zou, G. T., and Viberg, M., "Compensation for nonlinearity in a Hammerstein system using the coherence function with application to nonlinear acoustic echo cancellation," *IEEE Trans. Signal Process.*, vol. 55, pp. 5853–5858, Dec. 2007.

[110] Shimauchi, S. and Makino, S., "Stereo projection echo canceller with true echo path estimation," in *Proc. IEEE ICASSP*, pp. 3059–3062, May 1995.

[111] Shimauchi, S., Makino, S., Haneda, Y., Nakagawa, A., and Sakauchi, S., "A stereo echo canceller implemented using a stereo shaker and a duo-filter control system," in *Proc. IEEE ICASSP*, pp. 857–860, Mar. 1999.

[112] Shimauchi, S., Haneda, Y., and Kataoka, A., "Double-talk robust frequency domain echo cancellation algorithm with scalable nonlinear reference and error functions," in *Proc. IWAENC*, pp. 23–26, Sep. 2003.

[113] Shimauchi, S., Haneda, Y., and Kataoka, A., "A robust NLMS algorithm for acoustic echo cancellation," *Elect. Commun. Japan III: Fund. Elect. Sci.*, vol. 89, pp. 1–9, Aug. 2006.

[114] Shimauchi, S., Haneda, Y., and Kataoka, A., "Robust frequency domain acoustic echo cancellation filter employing normalized residual echo enhancement," *IEICE Trans. Fundamentals*, vol. E91-A, pp. 1347–1356, June 2008.

[115] Shynk, J. J., "Frequency-domain and multirate adaptive filtering," *IEEE Signal Process. Magazine*, vol. 9, pp. 14–37, Jan. 1992.

[116] Sondhi, M. M., "Adaptive echo cancellers," *Bell Syst. Tech. Journal*, vol. 46, pp. 497–511, Mar. 1967.

[117] Sondhi, M. M. and Mitra, D., "New results on the performance of a well-known class of adaptive filters," *Proc. IEEE*, vol. 64, pp. 1583–1597, Nov. 1976.

[118] Sondhi, M. M., Morgan, D. R., and Hall, J. L., "Stereophonic acoustic echo cancellation - an overview of the fundamental problem," *IEEE Signal Process. Letters*, vol. 2, pp. 148–151, Aug. 1995.

[119] Sondhi, M. M. and Presti, A. J., "A self-adaptive echo canceler," *Bell Syst. Tech. Journal*, vol. 45, pp. 1851–1854, Dec. 1966.

[120] Soo, J.-S. and Pang, K., "Multidelay block frequency domain adaptive filter," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 38, pp. 373–376, Feb. 1990.

[121] Stenger, A. and Kellermann, W., "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *Signal Process.*, vol. 80, pp. 1747–1760, Sep. 2000.

[122] Stenger, A. and Rabenstein, R., "An acoustic echo canceller with compensation of nonlinearities," in *Proc. EURASIP EUSIPCO*, pp. 969–972, Sep. 1998.

[123] Stenger, A., Trautmann, L., and Rabenstein, R., "Nonlinear acoustic echo cancellation with 2nd order adaptive Volterra filter," in *Proc. IEEE ICASSP*, vol. 2, pp. 877–880, Mar. 1999.

[124] Sugiyama, A., "A robust NLMS algorithm with a novel noise modeling based on stationary/nonstationary noise decomposition," in *Proc. IEEE ICASSP*, pp. 201–204, Apr. 2009.

[125] Sugiyama, A., Joncour, Y., and Hirano, A., "A stereo echo canceler with correct echo-path identification based on an input-sliding technique," *IEEE Trans. Signal Process.*, vol. 49, pp. 2577–2587, Nov. 2001.

[126] Valin, J.-M., "On adjusting the learning rate in frequency domain echo cancellation with double-talk," *IEEE Trans. Audio Speech Language Process.*, vol. 15, pp. 1030–1034, Mar. 2007.

[127] Valin, J.-M. and Collings, I. B., "A new robust frequency domain echo canceller with closed-loop learning rate adaptation," in *Proc. IEEE ICASSP*, vol. 1, pp. 93–96, Apr. 2007.

[128] Wada, T. S. and Juang, B.-H., "Enhancement of residual echo for improved acoustic echo cancellation," in *Proc. EURASIP EUSIPCO*, pp. 1620–1624, Sep. 2007.

[129] Wada, T. S. and Juang, B.-H., "Enhancement of residual echo for improved frequency-domain acoustic echo cancellation," in *Proc. IEEE WASPAA*, pp. 175–178, Oct. 2007.

[130] Wada, T. S. and Juang, B.-H., "Towards robust acoustic echo cancellation during double-talk and near-end background noise via enhancement of residual echo," in *Proc. IEEE ICASSP*, pp. 253–256, Apr. 2008.

[131] Wada, T. S. and Juang, B.-H., "Acoustic echo cancellation based on independent component analysis and integrated residual echo enhancement," in *Proc. IEEE WASPAA*, pp. 205–208, Oct. 2009.

[132] Wada, T. S. and Juang, B.-H., "Multi-channel acoustic echo cancellation based on residual echo enhancement with effective channel decorrelation via resampling," in *Proc. IWAENC*, Sep. 2010.

[133] Wada, T. S. and Juang, B.-H., "Enhancement of residual echo for robust acoustic echo cancellation," *IEEE Trans. Audio Speech Language Process.*, vol. 20, pp. 175–189, Jan. 2012.

[134] Wada, T. S., Juang, B.-H., and Sukkar, R. A., "Measurement of the effects of nonlinearities on the network-based acoustic echo cancellation," in *Proc. EURASIP EUSIPCO*, Sep. 2006.

[135] Wada, T. S., Miyabe, S., and Juang, B.-H., "Use of decorrelation procedure for source and echo suppression," in *Proc. IWAENC*, Sep. 2008, paper no. 9086.

[136] Wada, T. S., Robledo-Arununcio, E., Yue, G., and Juang, B.-H., "Immersive acoustic signal processing for intelligent collaboration,," in *Proc. WESPAC*, June 2006, paper no. 653.

[137] Wada, T. S., Wung, J., and Juang, B.-H., "Decorrelation by resampling in frequency domain for multi-channel acoustic echo cancellation based on residual echo enhancement," in *Proc. IEEE WASPAA*, pp. 289–292, Oct. 2011.

[138] Widrow, B. and Hoff, Jr., M. E., "Adaptive switching circuits," *IRE Wescon Conf. Rec.*, pp. 96–104, 1960, Part 4.

[139] Widrow, B. and Stearns, S. D., *Adaptive Signal Processing*. Prentice Hall, 1985.

[140] Wightman, F. and Kistler, D., "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. America*, vol. 91, pp. 1648–1661, Mar. 1992.

[141] Wung, J., Wada, T. S., and Juang, B.-H., "Inter-channel decorrelation by subband resampling in frequency domain," in *Proc. IEEE ICASSP*, pp. 29–32, Mar. 2012.

[142] Wung, J., Wada, T. S., Juang, B.-H., Lee, B., Kalker, T., and Schafer, R., "System approach to residual echo suppression in robust hands-free teleconferencing," in *Proc. IEEE ICASSP*, pp. 445–448, May 2011.

[143] Wung, J., Wada, T. S., Juang, B.-H., Lee, B., Trott, M., and Schafer, R. W., "A system approach to acoustic echo cancellation in robust hands-free teleconferencing," in *Proc. IEEE WASPAA*, pp. 101–104, Oct. 2011.

[144] Yang, J.-M. and Sakai, H., "A robust ICA-based adaptive filter algorithm for system identification," *IEEE Trans. Circuits Syst. II: Express Briefs*, vol. 55, pp. 1259–1263, Dec. 2008.

[145] Ye, H. and Wu, B.-X., "A new double-talk detection algorithm based on the orthogonality theorem," *IEEE Trans. Commun.*, vol. 39, pp. 1542–1545, Nov. 1991.

[146] Zou, Y., Chan, S.-C., and Ng, T.-S., "Least mean M-estimate algorithms for robust adaptive filtering in impulse noise," *IEEE Trans. Circuits Syst. II: Analog and Digital Signal Proc.*, vol. 47, pp. 1564–1569, Dec. 2000.

VITA

Ted S. Wada was born in Tokyo, Japan, and moved to the United States during the summer of 1982. After earning B.S. and M.S. degrees in applied physics, statistics, and electrical and computer engineering from Columbia University in New York, NY, Georgia State University in Atlanta, GA, and Georgia Institute of Technology in Atlanta, GA, he is about to finalize the academic pursuit with Ph.D. degree in electrical and computer engineering from Georgia Tech. He spent the summer of 2003 at Avaya, Inc., in Basking Ridge, NJ, the summer of 2005 at Tellabs, Inc., in Naperville, IL, and the winter/spring of 2009/2010 at Li Creative Technologies, Inc., in Florham Park, NJ, while working on acoustic echo cancellation and speech enhancement problems. He has been with Broadcom, Inc., in Irvine, CA, since the summer of 2012. His research interests include speech and audio signal processing, statistical signal processing, and pattern recognition.